

AD-A104 387

FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OH  
RECENT SELECTED PAPERS OF NORTHWESTERN POLYTECHNICAL UNIVERSITY--ETC(U)

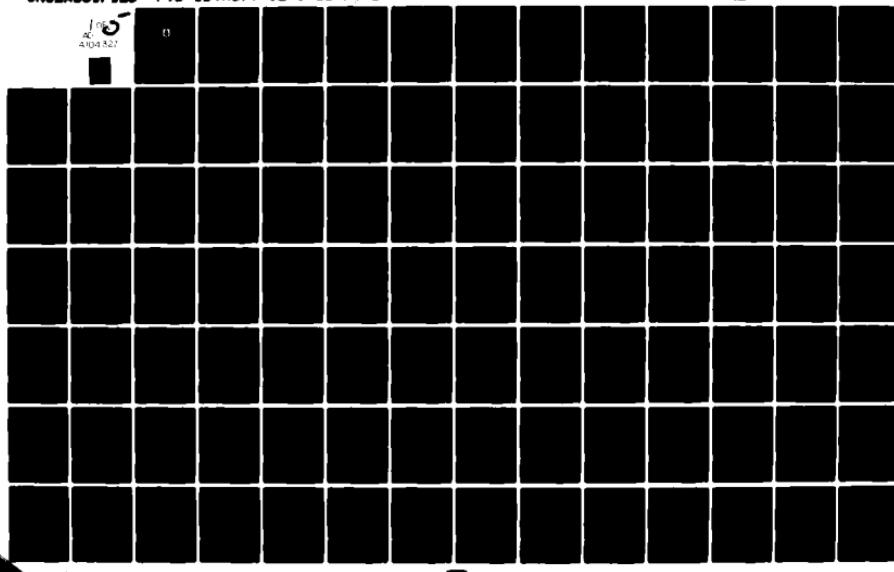
F/6 20/4

AUG 81

UNCLASSIFIED FTD-ID(RS)T-0259-81-PT-1

NL

100  
4104 327



DTIC FILE COPY

AD A104327

FTD-ID(RS)T-0259-81

2

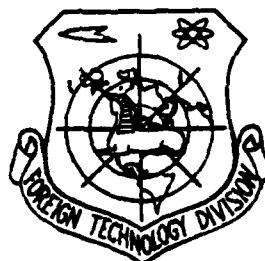
## FOREIGN TECHNOLOGY DIVISION



DTIC  
SEP 18 1981

E

RECENT SELECTED PAPERS OF  
NORTHWESTERN POLYTECHNICAL UNIVERSITY  
IN TWO PARTS. PART I. 1979



Approved for public release;  
distribution unlimited.

81 9 10

FTD-ID(RS)T-0259-81

## EDITED TRANSLATION

FTD-ID(RS)T-0259-81

20 August 1981

MICROFICHE NR: FTD-81-C-000754

RECENT SELECTED PAPERS OF NORTHWESTERN  
POLYTECHNICAL UNIVERSITY IN TWO PARTS.  
PART I. 1979.

English pages: 378

Source: Recent Selected Papers of Northwestern  
Polytechnical University in Two Parts, *Shanxi*,  
Part I. Xi'an-Shaaxi, 1979, pp. 1-3,  
1-4, 1-25, 27-35, 37-49, 51-105, 107-200,  
Pages 26, 36, 50, 106 missing in original  
foreign text.

Country of origin: China

Translated by: SCITRAN  
F33657-78-D-0619

Requester: FTD/TQTA

Approved for public release; distribution  
unlimited.

THIS TRANSLATION IS A RENDITION OF THE ORIGINAL FOREIGN TEXT WITHOUT ANY ANALYTICAL OR EDITORIAL COMMENT. STATEMENTS OR THEORIES ADVOCATED OR IMPLIED ARE THOSE OF THE SOURCE AND DO NOT NECESSARILY REFLECT THE POSITION OR OPINION OF THE FOREIGN TECHNOLOGY DIVISION.

PREPARED BY:

TRANSLATION DIVISION  
FOREIGN TECHNOLOGY DIVISION  
WP-AFB, OHIO.

FTD-ID(RS)T-0259-81

Date 20 Aug 1981

RECENT SELECTED PAPERS OF  
NORTHWESTERN POLYTECHNICAL UNIVERSITY  
IN TWO PARTS. PART I. 1979

Xi'an, Shaanxi  
The People's Republic of China

Accession For	
NTIS GENI	<input checked="" type="checkbox"/>
BUFILE	<input type="checkbox"/>
Library	<input type="checkbox"/>
Journal	<input type="checkbox"/>
Date	
1970	1970
Title	
Author	
Subject Codes	
Notes	
Date Received	
A	

## Foreword

In the present-day world, the promotion of international educational and scientific exchanges provides one of the important channels for scholars, scientists, engineers, and educators from different countries to become friends and to benefit from each other's experiences, thus contributing to the advancement of science and technology. For this reason, we have compiled "Recent Selected Papers of Northwestern Polytechnical University, in Two Parts, 1979." These papers are offered as a medium of scientific exchange and foreign friends are invited to oblige us with comments and suggestions.

We now have the great pleasure of availing ourselves of this opportunity to put before foreign scholars and friends some basic facts about the past and present of Northwestern polytechnical University (NPU) in order to acquaint them with NPU. We heartily welcome foreign scholars and professionals to visit NPU and deliver lectures. NPU, which is among the leading institutions of higher learning approved by the State Council, is a polytechnical university of aeronautical science and technology. Her campus is located at Xi'an (Sian), a city with thousands of years of history. In 1957, she was established by the amalgamation of the Northwestern Institute of Technology and the Xi'an Institute of Aeronautics. The Northwestern Institute of Technology had formed originally a part of the former Northwest Associated Universities established in 1938. Xi'an Institute of Aeronautics had been originally the East China Institute of Aeronautics, which had been established in 1952 by the amalgamation of the three Departments of Aeronautics in Shanghai Jiaotong (Chiaotung) University, Zhejiang (Chekiang) University, and Nanjing (Nanking) University (formerly Central University). In 1970, the Department of Aeronautics of the Haerbin (Harbin) Institute of Engineering moved to Xi'an to amalgamate with NPU and strengthened her personnel. At present, NPU has an instructional staff of over 1200, among which more than 250 are women, and over 900 are professors, associate professors, and lecturers. In 1979, there will be over 4000 undergraduate students and nearly 200 graduate students. Since the founding of the People's Republic of China in 1949, NPU has made great progress in education, research, and the construction of buildings and facilities. Over 17000 men and women, developed in an all-round way (morally, intellectually,

and physically) and trained in science and technology, have graduated from NPU. Many of them have risen to responsible positions in aeronautical and astronautical establishments. But as compared with advanced universities and institutes both within and outside China, NPU still lags considerably behind.

It is for well known reasons that the gap between China's and the advanced world scientific and technical level which had been narrowing widened again in the decade 1966-1976. The people of China is determined to attain at the end of this century the goal of "four modernizations", i.e., the socialist modernization of agriculture, industry, national defence, and science and technology. This requires arduous effort on our part to take effective measures for a vigorous upsurge in education. Therefore, we earnestly hope to exchange experiences concerning academic programs and scientific research with friendly foreign educators and scientists. With self-reliance as the underlying base, we will study conscientiously the advanced science and technology of foreign countries and exert every effort to turn them to useful account, gradually realizing the modernization of research facilities and teaching aids. We will do all these in order to transform our university into a center of excellence for education as well as for research as quickly as possible so that the potential of our university may be fully tapped for training the largest possible number of high-quality graduates.

The Academic and Research Council of  
Northwestern Polytechnical University

June 10, 1979

## Table of Abstracts

	<u>Page</u>
<i>Luo Shijun(Lo Shih-chun)*, Zheng Yuwen, Qian Hong, and Wang Dieqian</i>	
Finite Difference Computation of the Steady Transonic Potential Flow around Airplanes	5
<i>Lin Chaoqiang</i>	
Second Order Approximation Theory of an Arbitrary Aerofoil in Incompressible Potential Flow	33
<i>Lin Chaoqiang</i>	
Aerodynamic Calculations and Design of Subcritical Aerofoils	51
<i>Zhu Fangyuan, Zhou Xinhai, Liu Songling, and Fan Feida</i>	
An Aerodynamic Design Method for Transonic Axial Flow Compressor Stage	65
<i>Shen Huili and Chen Zhongqing</i>	
Analytical and Experimental Investigation of Performance of Supersonic Ejector Nozzle	87
<i>Liu Cikong and Zhao Jueliang</i>	
An Optimum Design Procedure of Total-Temperature Thermocouple Probes	111
<i>Wan Wei, Huang Jingxi, and Hu Shiming</i>	
A Synthesis Technique for Array Antennas of High Directivity and Low Sidelobe	132
104-105 <i>Dai Guanzhong</i>	
Model Method of State Estimation	155

\* Luo is the family name. Shijun is the given name; Prof. Luo's name was spelt Lo Shih-chun many years ago when he was a graduate student at the California Institute of Technology.

<i>Fu Hongzhi</i>	
Solidification Characteristics of Superalloys under Non-equilibrium Condition	191
<i>Zheng Xiulin, Qiao Shengru, and three 1978 graduates</i>	
The Plastic Deformation, Micro-crack Initiation, and Fatigue Crack Initiation Life of 30CrMnSiNi2A High Strength Martensite Steel	220
<i>Mei Shuoji</i>	
The Computation of Integral-Type Flexure Hinge Assembly of Dynamically Tuned Gyroscopes	240
<i>Shen Yunwen</i>	
A Theoretical Analysis of Involute Harmonic Gearing	272
<i>Chen Yizhou</i>	
A Method of Evaluation of the Torsional Rigidity and the Third stress Intensity Factor of Prismatical Bar of Rectangular Cross-section with Cracks	315
<i>Yang Qingxiong</i>	
The Automatic Matrix Force Method and Techniques for Handling More Complex Computations with Given Computer Capacity	338
<i>Ge Shoulian, Sun Can, Tang Xuanchun, and Ye Tianqi</i>	
Structural Analysis of Fuselages with Cutouts by Finite Element Method	359

## Summary

# Finite Difference Computation of the Steady Transonic Potential Flow around Airplanes

*Luo Shijun (Lo Shih-Chun), Zheng Yuwen,  
Qian Hong, and Wang Dieqian*

The velocity potential equation is approximated by assuming that the perturbation velocity component in the transverse plane of the body is much smaller than the undisturbed flow velocity, while the perturbation velocity component in the longitudinal direction may not be so. It is then solved with the mixed schemes of finite differences first used by Murman-Cole. The boundary conditions on wing (tails) and its wake are approximated similarly. The boundary condition on fuselage of arbitrary shape is satisfied by analytically continuing the potential field inside the fuselage. The boundary conditions at far field are calculated by following Klunker.

The finite difference equations for the velocity potential are solved by the line-relaxation method. The influence of the wing on the horizontal tail is computed with an approximate consideration of the deflection of the wing wake vortices. The pressure coefficient is calculated by the exact Bernoulli's equation.

Two numerical examples are included and the results agree fairly well with known wind tunnel test results:

(1) a wing-fuselage-horizontal tail-vertical tail combination of NACA TN 4041, Mach number  $M_\infty = 0.25$  and  $0.95$ ;

(2) the wing-fuselage combination of NASA TN D-830,  $M_\infty = 1.05$ , angle of attack  $\alpha = 2.2^\circ$ .

The meshes are  $38 \times 23 \times 17$  and  $25 \times 19 \times 19$  respectively. The iterative runs required for convergence are about 200 and 1500 respectively, while all iterations are initiated from zero perturbation potential.

In order to accelerate the convergence, example (2) is computed with the relaxation factor  $\omega = 1.7$  ( $M \leq 1$ ) and  $1.5$  ( $M > 1$ ).

FINITE DIFFERENCE COMPUTATION OF THE  
STEADY TRANSONIC POTENTIAL FLOW AROUND AIRPLANES

Luo Shijun (Lo Shih-Chun), Zheng Yuwen,  
Qian Hong and Wang Deiqian

Abstract

The velocity potential equation is approximated by assuming that the perturbation velocity component in the transverse plane of the bcdy is much smaller than the undisturbed flow velocity, while the perturbation velocity component in the longitudinal direction may not be so. It is then solved with the mixed schemes of finite differences first used by Murman-Cole. The boundary condition on wings(tails) and its wake are approximated similarly. The boundary condition on a fuselage of arbitrary shape is satisfied by analytically continuing the potential field inside the fuselage. The boundary conditions at far field are calculated by following Klunker.

The finite difference equations for the velocity potential are solved by the line-relaxation method. The influence of the wing on the horizontal tail is computed with an approximate consideration of the deflection of the wing wake vortices. The pressure coefficient is calculated by the exact Bernoulli equation.

Two numerical examples are included, and the results agree fairly well with known wind tunnel test results:

- (1) a wing-fuselage-horizontal tail-vertical tail combination of NACA TN 4041, Mach number  $M_\infty = 0.25$  and  $0.95$ ;
- (2) the wing-fuselage combination of NASA TN D-830,  $M = 1.05$ , angle of attack  $\alpha = 2.2^\circ$ .

The meshes are  $38 \times 23 \times 17$  and  $25 \times 19 \times 19$ , respectively. The iterative runs required for convergence are about 200 and 1500, respectively, while all iterations are initiated from zero perturbation potential.

In order to accelerate the convergence, example (2) is computed with the relaxation factor  $\omega = 1.7$  ( $M < 1$ ) and 1.5 ( $M > 1$ ).

## I. INTRODUCTION

The numerical computation of steady transonic potential flow has been well developed since Murman-Cole [1] proposed the mixed finite difference method. However, there is still a lack of practical computation methods for flows around bodies of complex combination such as airplanes, especially when the transonic free stream Mach number  $M_\infty \geq 1$ .

This study attempts to improve the accuracy of the differential equation for the perturbation velocity potential, the corresponding finite difference equation, the boundary condition and the formula for the pressure coefficient, without increasing the computation time significantly. It is assumed that the potential flow is perturbed except in the axial direction.

Due to the arbitrary profile of the combination body, the mesh for the finite difference calculation is directly taken from the orthogonal coordinate system of physical space, without any transformation of coordinates. To satisfy the boundary condition on the surface of a fuselage of arbitrary shape, a method of extending the velocity potential field analytically inside the fuselage is proposed here.

The computations for transonic flow field with both  $M_\infty < 1$  and  $M_\infty > 1$  cases are presented in the sample calculations. To accelerate the convergence speed of the iterations, a new investigation of the relaxation factor at certain locally supersonic points is made.

## II. THE VELOCITY POTENTIAL EQUATION AND THE BOUNDARY CONDITIONS

Taking the 0-x axis as the longitudinal axis of the airplane (Figure 1), and assuming the perturbation velocity components in the y and z directions are much less than the free-stream velocity, the perturbation velocity potential equation is obtained

$$(1 - M^2)\varphi_{ss} + \varphi_{yy} + \varphi_{zz} = 0 \quad (1)$$

where

$$1 - M^2 = \frac{1 - M_\infty^2 - \frac{\gamma + 1}{q_\infty} M_\infty^2 \varphi_s - \frac{\gamma + 1}{2q_\infty^2} M_\infty^2 \varphi_s^2}{1 - \frac{\gamma - 1}{q_\infty} M_\infty^2 \varphi_s - \frac{\gamma - 1}{2q_\infty^2} M_\infty^2 \varphi_s^2} \quad (2)$$

$M_\infty$  and  $q_\infty$  are the Mach number and velocity of the free-stream, respectively,

$M$  is the local Mach number,  
and  $\gamma$  is the adiabatic exponent of air.

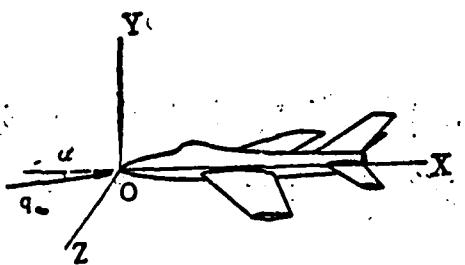


Figure 1

The exact relation is used for the boundary condition on the fuselage, that is, the following must be satisfied on the body surface

$$(q_\infty \cos \alpha + \varphi_\infty) F_x + (q_\infty \sin \alpha + \varphi_\infty) F_y + \varphi_\infty F_z = 0 \quad (3)$$

where  $F(x, y, z) = 0$  is the equation of the surface of the fuselage,  $\alpha$  is the angle of attack, defined as the angle between the free-stream velocity and the  $x$ -axis.

The following approximate relations are applied for the body surface of the wings, setting  $y = y_w$  for the wing surface:

$$\varphi_y(x, y_w + 0, z) = [q_\infty \cos \alpha + \varphi_\infty(x, y_w + 0, z)] \frac{\partial y_w}{\partial x} - q_\infty \sin \alpha \quad (4a)$$

$$\varphi_y(x, y_w - 0, z) = [q_\infty \cos \alpha + \varphi_\infty(x, y_w - 0, z)] \frac{\partial y_w}{\partial x} - q_\infty \sin \alpha \quad (4b)$$

where  $y_{wu}$  and  $y_{wl}$  are the  $y$  coordinates of the upper and lower surfaces of the wing respectively.

The approximate expressions of the Kutta condition for the wing vortex are given below, assuming on the wing vortex,  $y = y_w$ .

$$\varphi(x, y_w + 0, z) - \varphi(x, y_w - 0, z) = \varphi(x_{wt}, y_w + 0, z) - \varphi(x_{wt}, y_w - 0, z) \quad (5)$$

$$\varphi_y(x, y_w + 0, z) = \varphi_y(x, y_w - 0, z) \quad (5)$$

where  $x_{wt}$  is the  $x$  coordinate of the trailing edge of the wing.

The boundary conditions for the horizontal tail body surface and the corresponding vortex are the same as those for the wing.

The boundary condition at the body surface of the vertical tail is

$$\varphi_s(x, y, 0) = [q_\infty \cos \alpha + \varphi_s(x, y, 0)] \frac{\partial z_s}{\partial x} \quad (7)$$

where  $z_s$  is the  $z$  coordinate of the surface of the vertical tail.

The boundary condition at infinity is represented by the far field boundary condition. According to [2],

$$\varphi = \varphi_f + \varphi_w + \varphi_h + \varphi_{wf} + \varphi_{hf} \quad (8)$$

where  $\varphi_f$  is the far field contribution of the fuselage alone produced by the angle of attack,

$\varphi_w$ ,  $\varphi_h$  are the far field contributions of the wing and horizontal tail produced by the lift forces,

$\varphi_{wf}$ ,  $\varphi_{hf}$  are the additional far field contributions produced by the fuselage due to the lift interference of the wing and horizontal tail.

$$(1) \quad M_\infty < 1$$

$$\varphi_f = R^2(x) q_\infty \alpha \frac{y}{y^2 + z^2} \quad (9)$$

where  $R(x) = 1/4$  (the maximum height plus the maximum width of the fuselage at the location  $x$ ).

At the frontal far field boundary  $x = x_1$ ,

$$\varphi_w = \varphi_h = \varphi_{wf} = \varphi_{hf} = 0 \quad (10)$$

At the rear far field boundary  $x = x_2$ ,

$$\varphi_w = \begin{cases} \frac{\Delta\varphi_w(z)}{2}, & y = y_w + 0 \\ -\frac{\Delta\varphi_w(z)}{2}, & y = y_w - 0 \end{cases} \quad \left\{ \begin{array}{l} z_f \leq z \leq \frac{l}{2} \\ \frac{y - y_w}{2\pi} \int_{z_f}^{l/2} \left[ \frac{1}{(y - y_w)^2 + (z - \zeta)^2} + \frac{1}{(y - y_w)^2 + (z + \zeta)^2} \right] \Delta\varphi_w(\zeta) d\zeta \end{array} \right. \quad (11)$$

elsewhere

$$\varphi_{wf} = \frac{R^2(x)y}{\pi(y^2 + z^2)} \int_{z_f}^{l/2} \frac{\Delta\varphi_w(\zeta)}{\zeta^2} d\zeta \quad (12)$$

where  $\Delta\varphi_w(z) = \varphi(x_w, y_w + 0, z) - \varphi(x_w, y_w - 0, z)$ ,

$z_f$  is the z coordinate of the joint between the wing and fuselage,

$l$  is the span of the wing

At the upper, the lower and the left far field boundaries

$$\varphi_w = \frac{y - y_w}{2\pi[(y - y_w)^2 + z^2]} \left\{ 1 + \frac{x - x_w}{[(x - x_w)^2 + m^2(y - y_w)^2 + m^2z^2]^{1/2}} \right\} \int_{z_f}^{l/2} \Delta\varphi_w(\zeta) d\zeta \quad (13)$$

$$\varphi_{wf} = \frac{R^2(x)y}{2\pi(y^2 + z^2)} \int_{z_f}^{l/2} \frac{\Delta\varphi_w(\zeta)}{\zeta^2} d\zeta \cdot \begin{cases} \left[ 1 - \frac{x - x_w}{x_1 - x_w} \right], & x - x_w < 0 \\ \left[ 1 + \frac{x - x_w}{x_1 - x_w} \right], & x - x_w > 0 \end{cases} \quad (14)$$

where  $m = \sqrt{1 - M_\infty^2}$ ,

$x_w$  is the x coordinate of the mid-point of the mean aero-dynamic chord of the wing.

(2)  $M_\infty \geq 1$

The  $\phi_f$  expression at the frontal and rear far field is the same as before, but  $R(x)$  is converted into  $R(x - m\sqrt{y^2 + z^2})$ . At the upper, lower and left far field boundaries,

$$\varphi_u = \begin{cases} 0 & , \quad x - x_w \leq m\sqrt{(y - y_w)^2 + z^2} \\ \frac{y - y_w}{x\lceil(y - y_w)^2 + z^2\rceil} \frac{x - x_w}{[(x - x_w)^2 - m^2(y - y_w)^2 - m^2z^2]^{1/2}} \int_{-1}^{1/2} \Delta\varphi_w(\zeta) d\zeta, & x - x_w > m\sqrt{(y - y_w)^2 + z^2} \end{cases} \quad (15)$$

$$\varphi_{ul} = \begin{cases} 0 & , \quad x - x_w \leq m\sqrt{(y - y_w)^2 + z^2} \\ \frac{R^2(x - m\sqrt{y^2 + z^2})y}{2\pi(y^2 + z^2)} \left[ 1 + \frac{x - x_w - m\sqrt{(y - y_w)^2 + z^2}}{x_2 - x_w - m\sqrt{(y - y_w)^2 + z^2}} \right] \int_{-1}^{1/2} \frac{\Delta\varphi_w(\zeta)}{\zeta^2} d\zeta, & x - x_w > m\sqrt{(y - y_w)^2 + z^2} \end{cases} \quad (16)$$

$\phi_h$ ,  $\phi_{hf}$  are similar to  $\phi_w$ ,  $\phi_{wf}$ , except that the wing quantities in the expressions must be replaced by the corresponding quantities of the horizontal tail. Notice that the expression for  $\Delta\phi_h$  when  $y_w = y_h$  is different from that when  $y_w \neq y_h$ .

$$\Delta\varphi_h(z) = \begin{cases} \varphi(x_h, y_h + 0, z) - \varphi(x_h, y_h - 0, z), & y_w \neq y_h \\ \varphi(x_h, y_h + 0, z) - \varphi(x_h, y_h - 0, z) - \Delta\varphi_w(z), & y_w = y_h \end{cases} \quad (17)$$

### III. THE FINITE DIFFERENCE SCHEME

Due to the nature of differential equation (1), the mixed finite difference scheme is utilized for the derivatives in the  $x$  direction - that is, for locally subsonic ( $M < 1$ ) points, the central finite difference form is used.

$$\varphi_x = \frac{\varphi_{i+1,j,k}\Delta x_{i-1}^2 + \varphi_{i,j,k}(\Delta x_i^2 - \Delta x_{i-1}^2) - \varphi_{i-1,j,k}\Delta x_i^2}{\Delta x_{i-1}\Delta x_i(\Delta x_{i-1} + \Delta x_i)} + O(\Delta x^3) \quad (18)$$

$$\varphi_{xx} = 2 \frac{\varphi_{i+1,j,k}\Delta x_{i-1} - \varphi_{i,j,k}(\Delta x_{i-1} + \Delta x_i) + \varphi_{i-1,j,k}\Delta x_i}{\Delta x_{i-1}\Delta x_i(\Delta x_{i-1} + \Delta x_i)} + O(\Delta x) \quad (19)$$

and for locally supersonic ( $M \gg 1$ ) points, the backward finite difference form is used:

$$\phi_s = \frac{\phi_{i,j,k}[(\Delta x_{i-2} + \Delta x_{i-1})^2 - \Delta x_{i-1}^2] - \phi_{i-1,j,k}(\Delta x_{i-2} + \Delta x_{i-1})^2 + \phi_{i-2,j,k}\Delta x_{i-1}^2}{\Delta x_{i-2}\Delta x_{i-1}(\Delta x_{i-2} + \Delta x_{i-1})} \quad (20)$$

$$\phi_{ss} = 2 \frac{\phi_{i,j,k}\Delta x_{i-2} - \phi_{i-1,j,k}(\Delta x_{i-2} + \Delta x_{i-1}) + \phi_{i-2,j,k}\Delta x_{i-1}}{\Delta x_{i-2}\Delta x_{i-1}(\Delta x_{i-2} + \Delta x_{i-1})} + O(\Delta x) \quad (21)$$

where  $i, j, k$  are the mesh indices in the  $x, y, z$  directions, respectively, and  $\Delta x_i = x_{i+1} - x_i$

The central finite difference form is always used for the derivatives in the  $y$  and  $z$  directions, similar to expressions (18) and (19). Hence implicit finite difference equations are formed at locally supersonic or sonic points, ensuring stability in the course of finite difference calculations.

The rules for determining local velocities are listed in Table 1.

Table 1

$1 - M_c^2$	$1 - M_s^2$	Result
$\geq 0$	No effect	Subsonic points
$< 0$	$< 0$	Supersonic points
$< 0$	$\geq 0$	Sonic points

In the table,  $1 - M_c^2$  and  $1 - M_s^2$  express  $1 - M^2$ , the  $\phi_x$  of which are calculated by the central finite difference form and the backward finite difference form, respectively. At the sonic point,  $M = 1$ , equation (1) becomes

$$\varphi_{yy} + \varphi_{ss} = 0$$

(22)

According to Table 1, the finite difference form of equation (1) is non-conservative at the shock wave. The shock intensities obtained are weaker than those obtained by the exact shock wave relation, but agree quite well with experimental results [3]. Hence in the present study, the non-conservative finite difference equations given above are applied.

#### IV. THE INTRODUCTION OF BOUNDARY CONDITIONS

For a better simulation of transonic flow, it is required that both the boundary conditions and the velocity potential equation must be satisfied at the body surface and at the vortex sheet.

At the wing surface (or horizontal tail surface) and at the upper surface of its vortex sheet, the boundary conditions are introduced into the velocity potential equation through the following finite difference form of  $\phi_{yy}$ :

$$(\varphi_{yy})_{i,j+0,k} = \frac{2}{\Delta y_j} \left[ \frac{\varphi_{i,j+1,k} - \varphi_{i,j-1,k}}{\Delta y_j} - (\varphi_y)_{i,j+0,k} \right] + O(\Delta y) \quad (23)$$

where  $j + 0$  indicates the wing surface and the upper surface of the vortex sheet,

$(\phi_y)_{i,j+0,k}$  at the wing surface is substituted by boundary condition (4a), in which  $(\phi_x)_{i,j+0,k}$  has to be calculated by the mixed finite difference scheme (18) and (20). At the vortex sheet  $(\phi_y)_{i,j+0,k}$  is an unknown quantity.

At the mesh point A (i, j, k) of the upper surface of the fuselage (Figure 2), Equation (23) is also used for  $\phi_{yy}$ , in which  $(\phi_y)_{i, j, k}$  is calculated by boundary condition (3), that is

$$- \frac{[q_0 \cos \alpha + (\phi_s)_{i, j, k}] F_s + q_0 \sin \alpha F_y + (\phi_z)_{i, j, k} F_z}{F_s} \quad (24)$$

In the above equation, the mixed finite difference form is used for  $(\phi_x)_{i, j, k}$ , while the central finite difference form is used for  $(\phi_z)_{i, j, k}$ . In the course of calculating  $(\phi_z)_{i, j, k}$ , the mesh point B' inside the fuselage is reached (Figure 2). The mesh points inside the fuselage are outside the flow field, but it is assumed that there is also a perturbation velocity potential. It is the analytical extension of the perturbation velocity potential. Generally,  $(\phi_x)_{i, j, k}$  must be computed by means of the extended values too.

At the fuselage, wing surface and the lower surface of its vortex sheet, the boundary conditions can be inserted into the velocity potential equation similarly.

At the mesh point  $A_1$  (i, j, k) at the left of the fuselage surface (Figure 2), the boundary condition is introduced into the velocity potential equation with the following finite difference form of  $\phi_{zz}$ :

$$(\phi_{zz})_{i, j, k} = \frac{2}{\Delta z_k} \left[ \frac{\phi_{i, j, k+1} - \phi_{i, j, k}}{\Delta z_k} - (\phi_z)_{i, j, k} \right] + O(\Delta z) \quad (25)$$

where  $(\phi_z)_{i, j, k}$  is calculated by boundary condition (3), that is,

$$(\phi_z)_{i, j, k} = - \frac{[q_0 \cos \alpha + (\phi_s)_{i, j, k}] F_s + [q_0 \sin \alpha + (\phi_y)_{i, j, k}] F_y}{F_s} \quad (26)$$

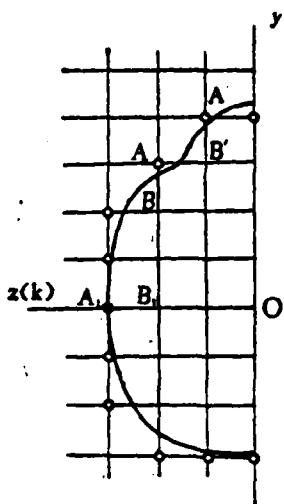


Figure 2

in which the mixed finite difference form is used for  $(\phi_x)_{i,j,k}$  while the central difference form is used for  $(\phi_y)_{i,j,k}$ . Generally the extended values are used. If the mesh points of the left surface of the fuselage fall on the wing or its vortex sheet, then  $(\phi_y)_{i,j,k}$  should be replaced by boundary condition (4) or the values at the vortex sheet.

At the plane of symmetry Oxy and the vertical tail surface, the finite difference equation can be set up similarly.

#### V. THE EXTENSION TOWARD THE INTERNAL SPACE OF THE FUSELAGE

The imaginary perturbation velocity potential inside the fuselage is obtained from the boundary conditions at the body surface. The point  $A(i, j, k)$  in Figure 2 is the mesh point on the upper surface, the perturbation velocity at the point  $B(i, j-1, k)$  immediately inside is solved by the central finite difference form of  $(\phi_y)_{i,j,k}$ :

$$\phi_{i,j-1,k} = \frac{\phi_{i,j+1,k} \Delta y_{j-1}^2 + \phi_{i,j,k} (\Delta y_j^2 - \Delta y_{j-1}^2) - (\phi_y)_{i,j,k} \Delta y_{j-1} \Delta y_j (\Delta y_{j-1} + \Delta y_j) + O(\Delta y^4)}{\Delta y_j^2} \quad (27)$$

in which  $(\phi_y)_{i,j,k}$  is replaced by Equation (24).

The internal extension of the upper surface of the fuselage is initiated from the highest point, for instance, the point A' in Figure 2. For that point, both  $(\phi_x)_{i,j,k}$  and  $(\phi_z)_{i,j,k}$  could be calculated by the velocity potential values of the flow field. In this way, the imaginary perturbation velocity potential of point B' is determined. Since the perturbation velocity potential at B' is now known, the perturbation velocity potential at point B can be determined similarly and so forth. The internal extension at the lower surface of the fuselage can be done in the same manner.

Consider the mesh point  $A_1(i, j, k)$  at the left surface of the fuselage in Figure 2. The imaginary perturbation velocity of its internal point  $B_1(i, j, k-1)$  is obtained by the central difference form of  $(\phi_z)_{i,j,k}$ :

$$\varphi_{i,j,k-1} = \frac{\varphi_{i,j,k+1} \Delta z_k^2 + \varphi_{i,j,k} (\Delta z_k^2 - \Delta z_{k-1}^2) - (\varphi_{i,j,k} \Delta z_{k-1} \Delta z_k (\Delta z_{k-1} + \Delta z_k))}{\Delta z_k^2} + O(\Delta z^3) \quad (28)$$

where  $(\phi_z)_{i,j,k}$  is calculated by Equation (26).

The internal extension from the left surface of the fuselage is initiated from the greatest width of the fuselage.

The extension toward the internal of the fuselage is made for two layers of mesh points. Hence it is required that at the diverging section of the fuselage, the difference in  $j$  between the adjacent mesh points of the body surface should not exceed 2, and the difference in  $k$  should not exceed 2 also. At the converging section of the fuselage, the difference in  $j$  between the adjacent mesh points of the body surface should not exceed 1, and the difference in  $k$  should not exceed 1 also.

When internal extensions are made from the upper, lower, and left surfaces of the fuselage, they might be overlapping each other. Hence the following rules are set: When the second layer points of the upper and lower extensions overlap with the first layer points, the second layer extension is omitted. When the points of the same layer of the upper and lower extensions overlap, the lower extension is omitted. When the points of the left extension overlap with the points of the upper and lower extensions, the left extension is omitted.

Due to the singularity of the perturbation velocity potential on the Ox axis, the analytical extension is not applicable when the extension points fall on the body axis Ox. Here, the average of the velocity potential of the upper and lower adjacent points is taken.

## VI. THE TREATMENT OF THE LEADING EDGE

The treatment at the leading edge is the same as the points on the wing surface. The leading edge is divided into an upper and a lower point, and the boundary conditions at the upper and the lower surfaces are introduced into the velocity potential equation respectively. Two problems arise:

- (1) The slopes of the upper and lower wing surfaces at the blunt leading edge,  $\frac{\partial y_u}{\partial x}, \frac{\partial y_l}{\partial x} \rightarrow \pm\infty$ . It is known that

$$\int_{x_l}^{x_t} \frac{\partial y_u}{\partial x} dx = y_u(x_t, z) - y_u(x_l, z) \quad (29)$$

where  $x_l$  and  $x_t$  are the x coordinates of the leading edge and the trailing edge.

The integral on the left hand side is calculated according to the finite difference mesh point values. The value of  $\frac{\partial y_1}{\partial x}$  at the leading edge found this way is finite. The value of  $\frac{\partial y_1}{\partial x}$  at the leading edge is determined in the same manner.

(2) The perturbation velocity potential at the leading edge is double-valued. Whenever the velocity potential at the leading edge is needed during the calculation of the velocity potential upstream of the leading edge (single valued), the mean value of the upper surface value and the lower surface value is taken. At the side edges of the wing and its wake vortex, the perturbation velocity potential is also double-valued and is treated in the same manner.

The vortex of the fuselage nose is treated as an internal point of the flow field.

#### VII. LINE RELAXATION AND ITERATION

The finite difference equation established above is non-linear. The coefficient  $1 - M^2$  of equation (1) (which is also the determinant for velocity) is calculated by the values of the velocity potential of the frontal field, that is, by Jacobian iteration. The set of linearized equations is solved by line-relaxed iteration along the y axis. The relaxation line proceeds forward according to the order  $k = 1, 2, 3, \dots$  on the planes normal to the x axis, and then proceeds from one plane to the other according to the order  $i = 1, 2, 3, \dots$ . During the line-relaxation process, the newest value of velocity potential is always taken, that is, Gauss-Seidel iteration is employed.

The unknown quantities of the relaxation along the y-axis are the perturbation velocity potential at the internal points

of the flow field, the body surface points, and the vortex points on the line, the upwash velocity ( $\phi_y$ )<sub>i, j, k</sub> at the vortex points. The internal points of the fuselage and the far field boundary points are excluded. These points are calculated by iterations of the extension formula and the far field formula, respectively.

Along each relaxation line in the y direction, the unknown quantities are arranged according to the order of j. Whenever the upwash velocity of the vortex sheet ( $\phi_y$ )<sub>i, j, k</sub> appears, it is placed between  $\phi_{i, j-0, k}$  and  $\phi_{i, j+0, k}$ . The coefficient matrices obtained are tri-diagonal. The set of equations is solved by the predictor-corrector method. To accelerate the convergence speed of the linearly relaxed iteration, the method of over-relaxation is employed. Whenever solutions of the equations are obtained, the following operation is done:

$$\varphi_{i,j,k}^{(n)} = \omega \bar{\varphi}_{i,j,k}^{(n)} + (1 - \omega) \varphi_{i,j,k}^{(n-1)} \quad (30)$$

where  $\bar{\varphi}_{i,j,k}^{(n)}$  is the n th iteration solution of the linearly relaxed equation set,  $\omega$  is the relaxation factor.

The undisturbed field is taken as the initial field of the iteration, that is,

$$\varphi_{i,j,k}^{(0)} = 0$$

Let  $|\varphi_{i,j,k}^{(n)} - \varphi_{i,j,k}^{(n-1)}|$  for all maximum values of i, j, k.

When the initial field is zero,  $\Delta\phi^{(1)}$  indicates the order of magnitude of the perturbation velocity potential. Hence

$$\frac{\Delta\phi^{(n)}}{\Delta\phi^{(1)}}$$

indicates the relative increment of the order of magnitude from the  $(n-1)^{th}$  iteration to the  $n^{th}$  iteration. For the engineering calculation, the convergence criterion is defined as

$$\frac{\Delta \varphi^{(n)}}{\Delta \varphi^{(n-1)}} \leq 10^{-3} \sim 10^{-6}$$

### VIII. THE DISTURBANCE OF THE WING ON THE HORIZONTAL TAIL

The free vortex of the wing undergoes a deflection after leaving the wing. To simulate the disturbance of the wing on the horizontal tail, the wing is traversed along the  $y$ -axis such that its vortex sheet is at the same level as the horizontal tail, as shown in Figure 3. The  $y$ -coordinate of the wing vortex at the mid-point of the mean aerodynamic chord ( $x_b, z_b$ ) of the exposed horizontal tail is defined as the  $y$ -coordinate of the entire vortex sheet, and is symbolized as  $y_v$ .

$$y_v = y_{wt}(z_b) + \alpha [x_b - x_{wt}(z_b)] + \frac{1}{q_\infty} \int_{x_{wt}(z_b)}^{x_b} \varphi_y(\xi, y_v, z_b) d\xi \quad (31)$$

where the value of  $\varphi_y$  at the location  $x_{wt}$  is determined by the velocity tangential to the airfoil.

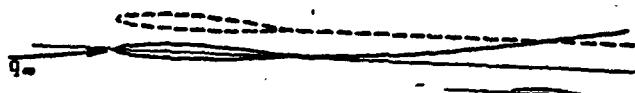


Figure 3

Replacing  $y_w$  with  $y_v$  and repeating the calculation of the entire flow field, the aerodynamic force on the horizontal tail of the body of combination is obtained. The aerodynamic force on the wing, however, should be taken from the calculated

results before the translation (that is, when it was at the position  $y_w$ ).

#### IX. THE EXPRESSION FOR PRESSURE COEFFICIENT

The exact form of the Bernoulli equation is applied to calculate the pressure coefficient:

$$\frac{p}{\rho M_\infty^2} = \frac{2}{\gamma M_\infty^2} \left\{ \left[ 1 - (\gamma - 1) M_\infty^2 \left( \frac{\varphi_x \cos \alpha}{q_\infty} + \frac{\varphi_y^2}{2q_\infty^2} + \frac{\varphi_y \sin \alpha}{q_\infty} + \frac{\varphi_z^2}{2q_\infty^2} + \frac{\varphi_z^2}{2q_\infty^2} \right) \right]^{\frac{2}{\gamma-1}} - 1 \right\} \quad (32)$$

In the above expression, all the derivatives are calculated by the central finite difference form (18). On the body surface, they are computed by means of the boundary conditions.

The lift coefficient  $c_y$ , drag coefficient  $c_x$  due to pressure difference and the pitching moment coefficient  $m_z$  are calculated from the pressure distribution on the body surface by the elliptical integral method.

#### X. SAMPLE CALCULATIONS

Example (1) is a combination of fuselage-wing-horizontal tail-vertical tail taken from reference [4]. For the sake of enabling the wing of the model to meet the requirement of calculating the disturbance on the horizontal tail without separating the wing from the fuselage after an upward translation, both the wing and horizontal tail are translated downwards for a distance of 1.51, under the condition of a fixed relative position between the two. It is assumed that the aerodynamic effect on the entire combination caused by this treatment is small. The end of the fuselage is connected to an infinitely long cylinder to simulate the effect of the supporting rod.

A  $38 \times 23 \times 17$  mesh system is employed. The origin is taken at the 41.5% point of the mean aerodynamic chord (18.48) of the wing. The far field boundaries are:

frontal, rear	$x_1 = -80.32, x_2 = 76.95$
upper, lower	$y = \pm 275$
left	$z = 49.4$

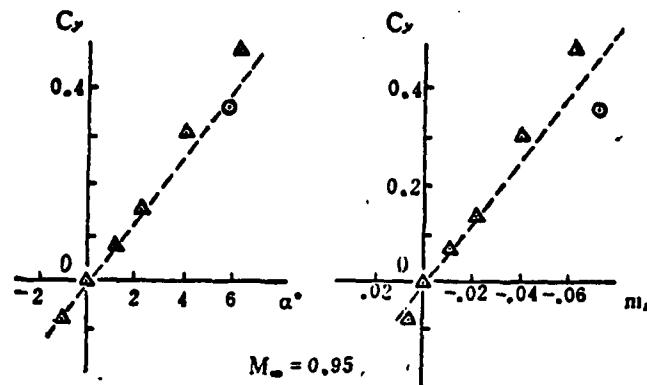
The exposed semi-span of the wing takes on 11 meshes, the root chord of the wing occupies 14 meshes, the semi-span of the horizontal tail takes on 7 meshes, its root chord spans 9 meshes, the span of the vertical tail takes on 4 meshes, its root chord occupies 13 meshes, the semi-width of the fuselage takes on 4 meshes, its length spans 33 meshes.

The cases of  $M_\infty = 0.25$  with an angle of attack of  $2.865^\circ$  and  $M_\infty = 0.95$  with an angle of attack of  $5.73^\circ$  are calculated. 197 and 150 iterations are required respectively to reach

$$\frac{\Delta \phi^{(n)}}{\Delta \phi^{(1)}} = 3 \times 10^{-4} \text{ and } 4 \times 10^{-3}$$

The two sets of computed results indicate that the upward translation along the y-axis of the wing vortex at the mid-point of the mean aerodynamic chord of the horizontal tail is less than half of the local step  $\Delta y = 0.91$ . Therefore, the effect due to the deflection of the wing vortex can be neglected. The comparison between the calculated results of the lift coefficient  $c_y$  and the pitching moment coefficient  $m_z$  and the corresponding experimental data [4] is shown in Figure 4.

Example (2) involves a fuselage-wing combination taken from [5]. The end of the fuselage is connected to an infinitely long cylindrical rod.



▲ experiment  
○ calculation

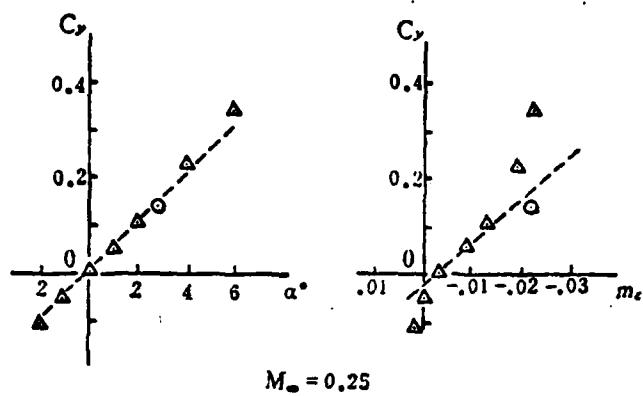


Figure 4

The mesh system is  $25 \times 19 \times 19$ . The origin is taken near the mid-point of the mean aerodynamic chord (30.23). The far field boundaries are:

frontal, rear  
upper, lower  
left

$x_1 = -200, x_2 = 200$   
 $y = \pm 250$   
 $z = 90$

The exposed semi-span of the wing occupies 11 meshes and its root chord takes on 12 meshes. The semi-width of the fuselage occupies 6 meshes and its length spans 21 meshes.

The case  $M_\infty = 1.05$ ,  $\alpha = 2.20^\circ$  is calculated, and the convergence is reached after 1500 iterations:

$$\frac{\Delta \varphi^{(n)}}{\Delta \varphi^{(1)}} = 10^{-4}$$

The calculated pressure distribution on the wing and fuselage is quite close to the experimental result [5], see Figure 5. Note that the calculated angle of attack of the fuselage is different from the experimental value. The calculated  $|\bar{p}|$  value is relatively low at the upper surface of the wing close to the leading edge.

For example (2), the relaxation factors are:

$$\omega = \begin{cases} 1.7, & M \leq 1 \\ 1.5, & M > 1 \end{cases}$$

The latter is selected according to the test computation introduced in section XI.

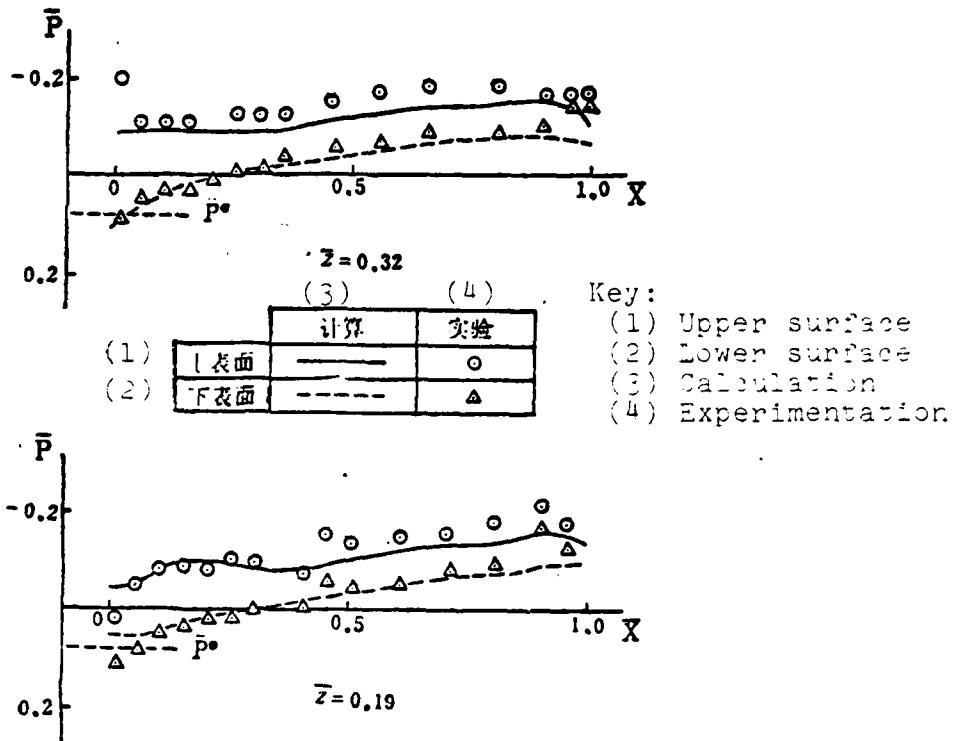


Figure 5a

## XI. THE TEST COMPUTATION FOR THE RELAXATION FACTOR

To reduce the computing time of the test run, only the exposed wing of the combination is considered during the trial. It is assumed that the relaxation factor selected for the wing alone is also appropriate for the wing-fuselage combination.

For the exposed wing of example (2), an  $18 \times 17 \times 13$  mesh system is taken. The origin is at the same point given above. The far field boundaries are:

frontal, rear	$x_1 = -65.38, x_2 = 73.68$
upper, lower	$y = \pm 256$
left	$z = 54.33$

The mesh position of the wing surface is the same as given

above. Different relaxation factors are employed to calculate for the case  $M_\infty = 1.05$ ,  $\alpha = 2.2^\circ$ .

The convergence of the relaxation iteration can be indicated by  $\Delta\phi^{(n)}$ . All the trial computations show that as  $n$  increases, the value of  $\Delta\phi^{(n)}$  fluctuates at the beginning. The fluctuation stops at a certain value of  $n$  and  $\Delta\phi^{(n)}$  becomes monotonically decreasing. The comparison between the converging process and the result is shown in Table 2. The convergence criterion is defined universally as

$$\Delta\phi^{(n)} < 10^{-8}$$

Table 2

$\omega$	$M \leq 1$	1.0	1.7	1.0	1.0	1.7	1.7	1.0
	$M > 1$	0.7	0.7	0.9	1.0	1.0	1.3	1.5
$\Delta\phi^{(1)}$	0.1390	0.1390	0.1862	0.2110	0.2110	0.3058	0.4192	
1. 停止振荡的 $n$	347	340	230	190	183	149	69	
2. 收敛的 $n$	465	459	365	338	335	218	153	
$c_v$	0.06293	0.06292	—	0.06324	0.0632	0.06553	0.06359	
$m_x$	-0.002729	-0.002727	—	-0.002741	-0.00278	-0.003455	-0.002802	
$c_s$	0.004533	0.004534	—	0.004496	0.00448	0.004341	0.004459	

Key: 1. When fluctuation stops,  $n =$  , 2. convergence  $n =$  .

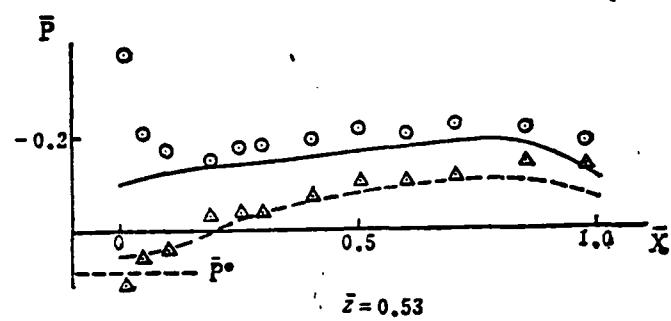
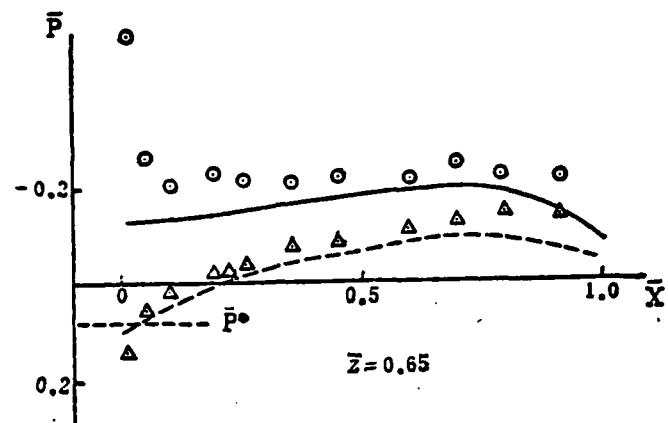


Figure 5b

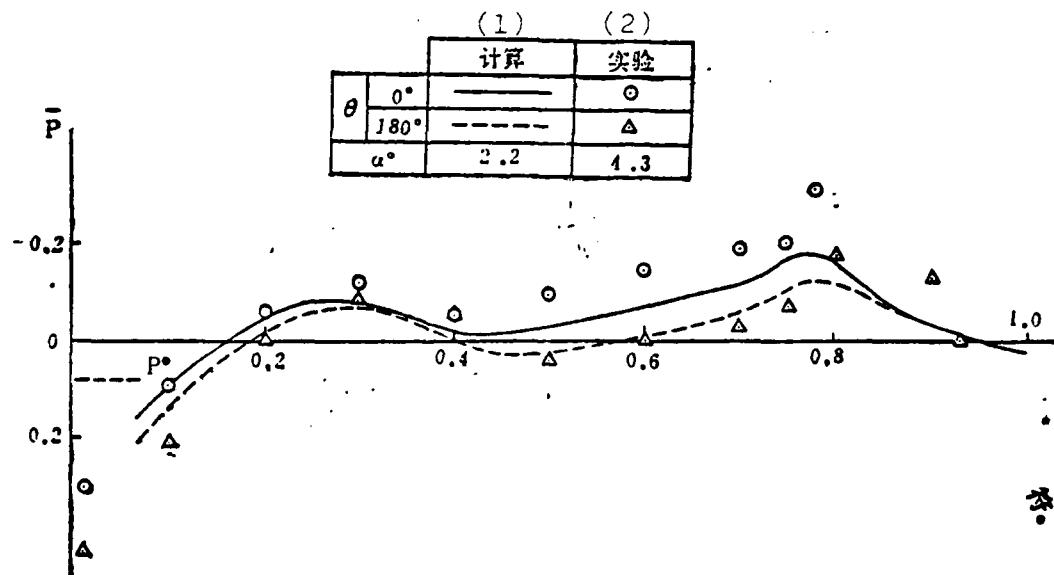


Figure 5c

Key: 1. Calculation, 2. Experiment.

As a verification of the accuracy of table 2,  $\omega = 1.7$  ( $M \leq 1$ ) and 1.0 ( $M > 1$ ) are taken and the calculations are repeated until  $\Delta\phi^{(n)} < 10^{-5}$ , for which  $n = 972$ . The aerodynamic coefficients are

$$c_y = 0.06371, \quad m_z = -0.002826, \quad c_x = 0.004450$$

which are close to the results given in the table above. The pressure distribution also agrees with that from the previous calculations. This indicates that the convergence criterion for the table above is accurate enough.

From Table 2, it is easily seen that for locally supersonic points the convergence is the fastest if  $\omega = 1.5$  is taken, while it is the slowest if  $\omega = 0.7$  is used. Hence, this sample calculation shows that convergence can be accelerated if a relaxation factor of value greater than 1 is taken for the locally supersonic points. This principle, however, is contradictory to reference [1].

### XII. CONCLUSION

This study presents a finite difference computation method for steady transonic perturbation potential flow, which is able to calculate airplanes consisting of an ordinary combination of fuselage, wing, horizontal tail and vertical tail. The pressure distribution and aerodynamic coefficients thus obtained agree with the experimental date.

In the sample calculations, transonic flows of both  $M_\infty < 1$  and  $M_\infty > 1$  are involved. When  $M_\infty > 1$ , the number of iterations increases dramatically. To accelerate the convergence, relaxation factor of values greater than 1 are taken for locally supersonic points.

### Appendix

The computations done for this study are executed on the 655 machine at the Hautung Research Institute for Computation Techniques, Shanghai.

### REFERENCES

- [1] Murman, E. M., and Cole, J. D., Calculation of Plane Steady Transonic Flows, AIAA J., 9(1971), 114-21.
- [2] Klunker, E. B., Contribution to Methods for Calculating the Flow About Thin Lifting Wings at Transonic Speeds—Analytic Expressions for Far Field, NASA TN D-6530 (1971).
- [3] Lomax, H., Bailey, F. R., and Ballhaus, W. F., On the Numerical Simulation of Three-Dimensional Transonic Flow with Application of the C-141 Wing, NASA TN D-6933 (1973).
- [4] Tinling, B. E., and Lopez, A. E., The Subsonic Static Aerodynamic Characteristics of an Airplane Model Having a Triangular Wing of Aspect Ratio 3.1—Effects of Horizontal-Tail Location and Size on the Longitudinal Characteristics, NACA TN 4041 (1957).
- [5] Swihart, J. M., and Foss, W. E. Jr., Transonic Loads Characteristics of 3%-Thick 60° Delta-Wing-Body Combination, NASA TN D-830 (1961)

## Summary

# Second Order Approximation Theory of an Arbitrary Aerofoil in Incompressible Potential Flow

*Lin Chaoqiang*

This paper presents a new method for calculating incompressible potential flow around an aerofoil. The region outside an arbitrary aerofoil is conformally transformed into the region outside a unit circle. The transformation functions are expanded into ascending power series and then only those terms up to and including the second order are retained. The resulting formulae are given analytically.

Although there exist several methods<sup>[1]-[8]</sup> for dealing with this classical problem, the present method manifests some important advantages.

The present method is nearly an exact solution from the engineering point of view, even for an aerofoil with thickness as much as 30% of the chord and with maximum camber as much as 10% of the chord.

Compared with the well-known Theodorsen's method, the computations involved can be carried out much more quickly and easily, without incurring any penalty on accuracy. Compared with Weber's method, the present method takes up nearly twice as much time for computation, but its accuracy is much better. In fact, it is shown by the author<sup>[11]</sup> that if only the first order terms are retained in the present method, the resulting formula is of the same form as that given by Weber's method. In other words, Weber's method has only a first order accuracy from the point of view of conformal transformation, and an additional term naturally and logically improves the accuracy of Weber's formula.

The present method, which is based on the smallness of thickness and maximum camber of the aerofoil (though it was mentioned above that the present method is none the less quite satisfactory even for those aerofoils having rather large relative thickness and large maximum relative camber), gives second order solutions which are uniformly valid throughout the flow field. Whereas in some other methods such as references [4][8], which are based on the smallness of perturbed velocity, the solutions break down in a region near the leading edge because the assumption of small velocity perturbation does not hold true there, and some techniques are needed as remedies.

SECOND ORDER APPROXIMATION THEORY OF AN  
ARBITRARY AEROFOIL IN INCOMPRESSIBLE POTENTIAL FLOW

Lin Chaoqiang

## SECOND ORDER APPROXIMATION THEORY OF AN ARBITRARY AEROFOIL IN INCOMPRESSIBLE POTENTIAL FLOW

### I. INTRODUCTION

This paper presents a second order approximation theory for calculating incompressible potential flow around the airfoil. The advantages of the method are simplicity of the calculation, high accuracy and its wide range of applications.

There exist several methods for dealing with this problem already, see references [1] - [8]. Theodorsen's method [1] can provide accurate solutions. Its key is to transform the region around the airfoil conformally into a region around a circle. The main purpose of the Theodorsen method is to obtain solutions by means of an implicit integral equation. Hence, multiple iterations are needed during the numerical computation, leading to an excessive computing time. The linearized thin airfoil theory [2] is simple to calculate but its accuracy is poor. Moreover, since the assumption of a small perturbation is not valid near the leading edge, the solution obtained for the entire flow field is not converging uniformly.

Goldstein [3] proposed an approximation theory on the basis of Theodorsen's method. For common airfoils, an approximation up to third order is usually required to obtain accurate results.

To improve the results from linearized thin airfoil theory, Lighthill [4] and Van Dyke, et. al. [8] introduced high order terms. Again, since the assumption of small perturbations is invalid near the leading edge, the solutions for higher order terms near the leading edge are found to diverge even faster than the first order terms. The authors of these two papers employed appropriate methods to modify the unreasonable results

near the leading edge.

Weber's method [5], [6] employs a conformal transformation of the region around the airfoil to a region around the branch cut, and obtains relatively simple calculation formulas after an appropriate simplifying assumption. In the Weber method, the airfoil is assumed to be thin and slightly cambered. But note that it does not demand a small angle of attack, while the methods presented by references [2], [4], [3] generally do prescribe a small angle of attack. That is why the Weber method is widely applied.

Similar to Weber's method, the method proposed by Spence et. al. [7] is also based on a conformal transformation of the region around the airfoil to a region around the branch cut. The thickness ratio of the airfoil is also assumed to be a minor factor. Spence et. al. expanded the conformal transformation relation with respect to the minor factors and preserved the terms up to the second order. Results with satisfactory accuracy were obtained. However, the Spence method is only limited to symmetrical airfoils.

The present method considers any arbitrary airfoil with camber, but it is assumed that the thickness ratio and the camber ratio of the airfoil are small. The region around the airfoil is first transformed conformally to a region around a unit circle. The transformation relation is then expanded with respect to a small factor and the terms up to the second order are preserved. Finally, explicit computation formulae are obtained. It is called the second order approximation theory.

The advantages of this method are:

- (1) Simplicity of calculation: On the TQ - 16 digital computer, the execution time required for the computation of each indivi-

dual airfoil is only 2.5 seconds.

(2) The results obtained are uniformly valid throughout the entire flow field. Modifications such as those given by reference [4] and [8] are not required.

(3) Wide range of applications. The theory itself demands small thickness and small camber of the airfoil. However, the sample calculations indicate that the accuracy of the computed results is still very satisfactory even for symmetric airfoils of 50% relative thickness and Zhukovskiy airfoil of 30% thickness ratio and 10% camber ratio. Also the angle of attack is not limited to being small.

By means of the common expression of the conformal transformation, this paper also presents simple formulae for the calculation of the zero-lift angle of attack, the slope of the lift coefficient, the pressure centre and the moment coefficient with respect to the pressure centre.

The Weber method is usually called non-linear theory. It has been proved in reference [11] that from the point of view of expanding the conformal transformation relation with respect to the small thickness and the small camber, the Weber method is actually an equivalence of the first order approximation theory. In that paper, we have pointed out that a term related to the camber distribution of the airfoil is practically missing in the Weber method. If this term is supplemented, the accuracy of Weber's method can be improved.

Since the Weber method is equivalent to the first order approximation theory of the conformal transformation, the accuracy of the method presented here is better than the Weber method. The sample calculations show that the accuracy of this method is apparently much better than that of the Weber

method, especially for those airfoils with camber.

## II. THE SECOND ORDER APPROXIMATION EXPANSION OF THE CONFORMAL TRANSFORMATION

Let the airfoil to be of unit length, with the leading edge located at the origin  $z = 0$  of the  $z$ -plane ( $z = x + iy$ ) and the trailing edge located at  $x = 1$ . Let the function

$$z = f(\zeta) = k\zeta + k_0 + \sum_{n=1}^{\infty} \frac{k_n}{\zeta^n} \quad (1)$$

transform conformally the region around the airfoil in the  $z$ -plane to a region around a unit circle with its center located at the origin in the  $\zeta$ -plane.  $k$  is a positive real number. According to Riemann's theorem, transformation (1) exists and is unique.

On the unit circle  $\zeta = e^{i\theta}$ . Consider the function  $i\{f(\zeta) - k\zeta - k_0\}$ . It approaches zero as  $\zeta \rightarrow \infty$ , and the real part on the unit circle is  $-y + k\sin \theta + k_{0R}$ , where  $y$  is the longitudinal coordinate of the airfoil surface,  $k_{0I}$  is the imaginary part of  $k_0$ . From Schwarz's formula:

$$i\{f(\zeta) - k\zeta - k_0\} = \frac{1}{2\pi i} \oint (-y + k\sin \theta + k_{0R}) \frac{\zeta + \zeta'}{\zeta - \zeta'} \cdot \frac{d\zeta'}{\zeta'},$$

that is

$$f(\zeta) - k\left(\zeta + \frac{1}{\zeta}\right) - k_{0R} = \frac{1}{2\pi} \oint y \frac{\zeta + \zeta'}{\zeta - \zeta'} \cdot \frac{d\zeta'}{\zeta'}, \quad (2)$$

where  $k_{0R}$  is the real part of  $k_0$ .

In equation (2), the airfoil surface coordinate  $y$  is an unknown function of  $\zeta$  before the conformal transformation is determined. Hence equation (2) cannot be treated as the

explicit relation for the transformation  $f(\zeta)$ .

The auxiliary coordinate  $\phi$  is now introduced. Let

$$x = \frac{1}{2}(1 + \cos\phi), \quad (3)$$

Assuming that the airfoil is thin and slightly cambered, then

$$y = \epsilon \cdot \psi(\phi), \quad (4)$$

where  $\epsilon$  is of the same order of the magnitude of the thickness ratio and the camber ratio of the airfoil, while  $\psi(\phi) = O(1)$ . As  $\epsilon \rightarrow 0$ , the airfoil reduces to a flat plate, and equation (1) reduces to the well known Zhukovskiy transformation

$$z = \frac{1}{2} + \frac{1}{4} \left( \xi + \frac{1}{\xi} \right)$$

Apparently we have  $\theta = \phi$ . Hence it can be assumed that

$$\theta = \phi + \epsilon \cdot F_1(\phi) + \epsilon^2 \cdot F_2(\phi) + \dots, \quad (5)$$

Also, the following assumptions are made correspondingly:

$$\begin{aligned} k &= k^{(0)} + \epsilon \cdot k^{(1)} + \epsilon^2 \cdot k^{(2)} + \dots, \\ k_{0K} &= k_{0K}^{(0)} + \epsilon \cdot k_{0K}^{(1)} + \epsilon^2 \cdot k_{0K}^{(2)} + \dots, \end{aligned} \quad \} \quad (6)$$

The real part on the unit circle in equation (2) is used. Making use of equations (5) and (6), both sides of equation (2) are expanded as power series of  $\epsilon$  up to the second order terms. Then equating the terms of the same power on both sides, the fundamental equations of the zeroth order, first order and second order approximations are obtained. They are:

(i) zeroth order approximation equation:

$$\frac{1}{2}(1+\cos\phi) - 2k^{(0)}\cos\phi - k_{0k}^{(0)} = 0, \quad (7)$$

(ii) first order approximation equation:

$$-2k^{(1)}\cos\phi + 2k^{(0)}F_1(\phi)\sin\phi - k_{0k}^{(1)} = \frac{1}{2\pi}P\int_0^{2\pi}\psi(\phi')\cot\frac{\phi-\phi'}{2}d\phi', \quad (8a)$$

(iii) second order approximation equation:

$$\begin{aligned} & -2k^{(2)}\cos\phi + 2k^{(1)}F_1(\phi)\sin\phi + 2k^{(0)}F_2(\phi)\sin\phi + k^{(0)}F_1^2(\phi)\cos\phi \\ & - k_{0k}^{(2)} = \frac{1}{2\pi}P\int_0^{2\pi}\frac{d\psi(\phi')}{d\phi'}[F_1(\phi) - F_1(\phi')]\cot\frac{\phi-\phi'}{2}d\phi', \end{aligned} \quad (9a)$$

The  $P$  preceding the integral signs in equations (8a) and (9a) means that the principal value of the Cauchy integral should be taken.

Another auxiliary plane  $\tau = \xi + i\eta$  is introduced. Its relation with the  $\zeta$ -plane is

$$\tau = k\left\{\xi + \frac{1}{\zeta} + 2\right\}, \quad (10)$$

It transforms the region around the unit circle of the  $\zeta$ -plane into a region above a branch cut  $[0, a]$  on the real axis ( $\xi$  axis) in the  $\tau$ -plane.  $a = 4k$ . Comparing equations (1) and (10), it can be seen that  $\left(\frac{dz}{d\tau}\right)_{\tau=0} = 1$ . If  $\zeta = e^{i\theta}$  is taken in equation (1), then

$$\xi = \frac{1}{2}a(1 + \cos\theta). \quad (11)$$

Applying  $a = 4k$  and equations (5) and (6), the above equation can be written as

$$\xi = x + \varepsilon \cdot f_1(\phi) + \varepsilon^2 \cdot f_2(\phi) + \dots, \quad (12)$$

where

$$f_1(\phi) = 2k^{(1)}(1 + \cos \phi) - 2k^{(0)}F_1(\phi) \sin \phi, \quad (13)$$

$$f_2(\phi) = 2k^{(2)}(1 + \cos \phi) - 2k^{(1)}F_1(\phi) \sin \phi - 2k^{(0)}F_2(\phi) \sin \phi - k^{(0)}F_1^2(\phi) \cos \phi. \quad (14)$$

Employing equations (13) and (14), equations (8a) and (9a) can be rewritten as

$$2k^{(1)} - k_{0k}^{(1)} - f_1(\phi) = \frac{1}{2\pi} P \int_0^{2\pi} \psi(\phi') \cot \frac{\phi - \phi'}{2} d\phi', \quad (8b)$$

$$2k^{(2)} - k_{0k}^{(2)} - f_2(\phi) = \frac{1}{2\pi} P \int_0^{2\pi} \frac{d\psi(\phi')}{d\phi'} [F_1(\phi) - F_1(\phi')] \cot \frac{\phi - \phi'}{2} d\phi' \quad (9b)$$

### III. THE SOLUTION OF THE ZEROTH ORDER APPROXIMATION EQUATION

Equation (7) is the zeroth order approximation equation. Its solution is obvious:

$$k^{(0)} = \frac{1}{4}, \quad k_{0k}^{(0)} = \frac{1}{2} \quad (15)$$

### IV. THE SOLUTION OF THE FIRST ORDER APPROXIMATION EQUATION

The first order approximation equation is given by equation (8a) or (8b).

To solve this equation, that is, to find the function  $f_1(\phi)$  or  $F_1(\phi)$ , it is assumed that the coordinate of the airfoil surface can be approximated by the following trigonometric polynomial:

$$\frac{y}{c} = \psi(\phi) = a_0 + \sum_{n=1}^{N-1} (a_n \cos n\phi + b_n \sin n\phi) + a_N \cos N\phi, \quad (16)$$

in which  $N$  is an even number. For the sample calculations presented by this paper,  $N = 18$ . The coefficients  $a_n$  ( $n = 0, 1, 2, \dots, N$ ) and  $b_n$  ( $n = 1, 2, \dots, N-1$ ) are determined by the corresponding value of the trigonometric polynomial (16) and the  $y$  value of the airfoil when  $\phi = \phi_p = p \pi/N$  ( $p = 0, 1, 2, \dots, 2N-1$ ). Following common practice, the coordinate of the airfoil is decomposed into two parts, that is, the thickness distribution and the camber distribution:

$$\left. \begin{array}{l} \text{thickness distribution } \psi_t(\phi) = \frac{1}{2} \{ \psi(\phi) - \psi(-\phi) \}, \\ \text{camber distribution } \psi_c(\phi) = \frac{1}{2} \{ \psi(\phi) + \psi(-\phi) \}, \end{array} \right\} \quad (17)$$

according to reference [7], the values of the coefficients in equation (16) are

$$\left. \begin{array}{l} a_0 = \frac{1}{N} \sum_{r=1}^{N-1} \psi_c(\phi_r), \quad a_N = \frac{1}{N} \sum_{r=1}^{N-1} (-1)^r \psi_c(\phi_r), \\ a_n = \frac{2}{N} \sum_{r=1}^{N-1} \psi_t(\phi_r) \cos n\phi_r, \\ b_n = \frac{2}{N} \sum_{r=1}^{N-1} \psi_t(\phi_r) \sin n\phi_r. \end{array} \right\} \quad (n = 1, 2, \dots, N-1) \quad (18)$$

Substituting equation (16) into (8b), the following equation is obtained:

$$2k^{(1)} - k_{0N}^{(1)} - f_1(\phi) = \sum_{n=1}^{N-1} (a_n \sin n\phi - b_n \cos n\phi), \quad (b_N = 0) \quad (19)$$

Substituting  $\phi = 0$  and  $\phi = \pi$  into the above equation separately and getting  $f_1(0) = 4 k^{(1)}$  and  $f_1(\pi) = 0$  from equation (13), a set of simultaneous equations to determine  $k^{(1)}$  and  $k_{0N}^{(1)}$  is obtained. Solving this set of equations, it is found that

$$k^{(1)} = \frac{1}{2} \sum_{n=1, 3, \dots, N-1} b_n, \quad k_{0N}^{(1)} = \sum_{n=2, 4, \dots, N} b_n \quad (20a)$$

Solving  $f_1(\phi)$  from equation (19) and then substituting it into (13), the solution for  $F_1(\phi)$  is obtained:

$$F_1(\phi) = \frac{2(k_{0u}^{(1)} + 2k^{(1)} \cos \phi)}{\sin \phi} + \frac{2}{\sin \phi} \sum_{n=1}^N (a_n \sin n\phi - b_n \cos n\phi),$$

Substituting (20a) into the above equation, it becomes:

$$F_1(\phi) = \sum_{n=1}^{N-1} (c_n \cos n\phi + d_n \sin n\phi), \quad (21)$$

where

$$\left. \begin{aligned} c_{N-1} &= 4a_N, \\ c_{N-2} &= 4a_{N-1}, \\ c_n &= 4a_{n+1} + c_{n+2} = 4(a_{n+1} + a_{n+2} + \dots), \\ (n &= N-3, N-4, \dots, 2, 1) \end{aligned} \right\} \quad (22a)$$

$$\left. \begin{aligned} d_{N-1} &= 0, \\ d_{N-2} &= 4b_{N-1}, \\ d_{N-3} &= 4b_{N-2}, \\ d_n &= 4b_{n+1} + d_{n+2} = 4(b_{n+1} + b_{n+2} + \dots), \\ (n &= N-4, N-5, \dots, 2, 1) \end{aligned} \right\} \quad (22b)$$

Let

$$d_0 = 2(b_1 + b_2 + \dots + b_{N-1}), \quad (23)$$

Then from equations (22b) and 23), equation (20a) can be written as:

$$k^{(1)} = \frac{1}{4}d_0, \quad k_{0u}^{(1)} = \frac{1}{4}d_1. \quad (20b)$$

Substituting (20b) into equation (19), the following equation is obtained:

$$f_1(\phi) = \frac{1}{2}d_0 - \frac{1}{4}d_1 - \sum_{n=1}^N (a_n \sin n\phi - b_n \cos n\phi) \quad (24)$$

## V. THE SOLUTION OF THE SECOND ORDER APPROXIMATION EQUATION

To solve the second order approximation basic equation (9b), the integral at the right hand side of equation (9b) must be found first. This can be done by substituting the  $\psi(\phi)$  of equation (16) and the  $F_1(\phi)$  of equation (21) into the integral, and then integrating term by term. The integral at the right hand side of equation (9b) is thus determined.

$$\begin{aligned} & \frac{1}{2\pi} P \int_0^{2\pi} \frac{d\psi(\phi')}{d\phi'} [F_1(\phi) - F_1(\phi')] \cot \frac{\phi - \phi'}{2} d\phi' \\ &= \sum_{m=1}^N \sum_{n=1}^{N-1} m(b_m c_n I_{m,n} - a_m d_n J_{m,n} + b_m d_n K_{m,n} - a_m c_n L_{m,n}), \end{aligned} \quad (25)$$

where

$$\begin{aligned} I_{m,n} &= \frac{1}{2\pi} P \int_0^{2\pi} \cos m\phi' (\cos n\phi - \cos n\phi') \cot \frac{\phi - \phi'}{2} d\phi' \\ &= \begin{cases} 0, & m \geq n \\ -\sin(n-m)\phi, & m < n \end{cases} \end{aligned} \quad (26)$$

$$\begin{aligned} J_{m,n} &= \frac{1}{2\pi} P \int_0^{2\pi} \sin m\phi' (\sin n\phi - \sin n\phi') \cot \frac{\phi - \phi'}{2} d\phi' \\ &= I_{m,n}, \end{aligned} \quad (27)$$

$$\begin{aligned} K_{m,n} &= \frac{1}{2\pi} P \int_0^{2\pi} \cos n\phi' (\sin n\phi - \sin n\phi') \cot \frac{\phi - \phi'}{2} d\phi' \\ &= \begin{cases} 0, & m > n \\ \frac{1}{2}, & m = n \\ \cos(n-m)\phi, & m < n \end{cases} \end{aligned} \quad (28)$$

$$\begin{aligned} L_{m,n} &= \frac{1}{2\pi} P \int_0^{2\pi} \sin m\phi' (\cos n\phi - \cos n\phi') \cot \frac{\phi - \phi'}{2} d\phi' \\ &= -K_{m,n}. \end{aligned} \quad (29)$$

Substituting equations (26) - (29) into (25), then into equation (9b), and manipulating, the following equation is obtained:

$$2k^{(2)} - k_{0N}^{(2)} - f_2(\phi) = \sum_{n=0}^{N-2} (e_n \cos n\phi + f_n \sin n\phi), \quad (30)$$

in which

$$\left. \begin{aligned} e_0 &= \frac{1}{2} \sum_{m=1}^{N-1} m(b_m d_m + a_m c_m), \\ f_0 &= \frac{1}{2} \sum_{m=1}^{N-1} m(a_m d_m - b_m c_m), \\ e_n &= \sum_{m=1}^{N-n-1} m(b_m d_{m+n} + a_m c_{m+n}), \\ f_n &= \sum_{m=1}^{N-n-1} m(a_m d_{m+n} - b_m c_{m+n}). \end{aligned} \right\} (n=1, 2, \dots, N-2) \quad (31)$$

From equation (14), the following expressions can be obtained:

$$f_2(0) = 4k^{(2)} - \frac{1}{4} F_1^2(0), \quad f_2(\pi) = \frac{1}{4} F_1^2(\pi) \quad (32)$$

Employing equation (32) and substituting  $\phi=0$  and  $\phi=\pi$  into equation (30) respectively, the set of simultaneous equations determining  $k^{(2)}$  and  $k_{0N}^{(2)}$  is obtained. Solving this set of equations, we have

$$\left. \begin{aligned} k^{(2)} &= \frac{1}{16} [F_1^2(0) + F_1^2(\pi)] - \frac{1}{2} \sum_{n=1, 3, \dots, N-2} e_n, \\ k_{0N}^{(2)} &= \frac{1}{8} [F_1^2(0) - F_1^2(\pi)] - \sum_{n=2, 4, \dots, N-2} e_n. \end{aligned} \right\} \quad (33)$$

According to equation (21), it is apparent that in the above equations,

$$F_1(0) = \sum_{n=1}^{N-1} e_n, \quad F_1(\pi) = \sum_{n=1}^{N-1} (-1)^n c_n \quad (34)$$

Substituting equation (33) into equation (30), the following expression can be obtained:

$$f_2(\phi) = \frac{1}{4} F_1^2(\pi) + \sum_{n=0}^{N-2} (-1)^n e_n - \sum_{n=0}^{N-2} (e_n \cos n\phi + f_n \sin n\phi) \quad (35)$$

## VI. VELOCITY DISTRIBUTION ON THE AIRFOIL SURFACE

Once the conformal transformation between the region around the airfoil and the region around the unit circle is found, the velocity distribution on the airfoil surface can be determined without much difficulty.

Firstly the  $\theta$  values on the unit circle in the  $\zeta$ -plane corresponding to the points on the airfoil should be determined. The procedure:  $f_1(\phi)$  and  $f_2(\phi)$  are solved from equations (24) and (35) and are substituted into equation (12) to solve for  $\xi$ . Substituting  $\xi$  into equation (11) and employing the inverse trigonometric relations, the value of  $\theta$  is found. (Note that in equation (11),  $a = 4k$ , and the value of  $k$  can be determined by equations (6), (15), (20b) and (33)). The value of the inverse trigonometric function should be determined by the sign of  $\frac{d\xi}{d\phi}$ : when  $\frac{d\xi}{d\phi} \leq 0$ ,  $\theta$  should assume values in the I, II quadrants, whereas when  $\frac{d\xi}{d\phi} > 0$ ,  $\theta$  should fall in the III, IV quadrants. The formula for calculating  $\frac{d\xi}{d\phi}$  can be obtained by substituting equations (3), (24) and (35) into equation (12) and differentiating once.

From equation (1), it is found that  $\left(\frac{dz}{d\xi}\right)_{\xi=0} = k$ .

Hence if the free stream velocity in the physical plane is  $V_\infty$ , then the free stream velocity in the  $\xi$ -plane is  $kV_\infty$ . Let  $\alpha$  be the angle of attack. According to the solution of the flow around a circular cylinder, the velocity on the unit circle in the  $\xi$ -plane is (without loss of generality,  $V_\infty = 1$  is taken):

$$V_\xi = 4k \sin \frac{\theta - \alpha_0}{2} \cos \left( \alpha - \frac{\theta + \alpha_0}{2} \right) \quad (36)$$

where  $\alpha_0$  is the  $\theta$  value at the trailing edge, that is, the zero lift angle of attack of the airfoil. The velocity on the airfoil surface in the physical plane is

$$V = V_\xi / \left| \frac{dz}{d\xi} \right| \quad (37)$$

where

$$\left| \frac{dz}{d\xi} \right| = \left| \frac{dz}{d\phi} \right| \cdot \left| \frac{d\xi}{d\theta} \right| / \left| \frac{d\xi}{d\phi} \right| \quad (38)$$

In equation (38), the evaluation of  $\frac{d\xi}{d\phi}$  has been accounted for before. From equation (11),

$$\frac{d\xi}{d\theta} = -\frac{1}{2} \alpha \sin \theta \quad (39)$$

Also,

$$\left| \frac{dz}{d\phi} \right| = \sqrt{\left( \frac{dx}{d\phi} \right)^2 + \left( \frac{dy}{d\phi} \right)^2} \quad (40)$$

In equation (40), the expressions for  $\frac{dx}{d\phi}$  and  $\frac{dy}{d\phi}$  are found by differentiating equations (3) and (16) with respect to  $\phi$ .

In the above manipulations, an exceptional case must also be considered, that is, when  $\sin \theta = 0$  or its absolute value is very small, the absolute values of  $\frac{d\xi}{d\theta}$  and  $\frac{d\xi}{d\phi}$  are also zero or very small. Hence equation (38) is indeterminate. For such a situation, the L'Hospital rule must be applied, and the following limit is obtained:

$$\lim_{\sin \theta \rightarrow 0} \left| \frac{\frac{d\xi}{d\theta}}{\frac{d\xi}{d\phi}} \right| = \sqrt{\frac{2}{a}} \left| \frac{d^2\xi}{d\phi^2} \right| \quad (41)$$

The value of  $\left| \frac{d^2\xi}{d\phi^2} \right|$  in the above equation can be obtained by differentiating twice the  $\xi-\phi$  functional relation found previously.

## VII. SOME AERODYNAMIC CHARACTERISTICS OF THE AIRFOIL

The zero lift angle of attack  $\alpha_0$  is exactly the angle corresponding to the trailing edge of the airfoil. It has been discussed already. According to the theory of incompressible potential flow around an airfoil (for example, reference [10], only the first three coefficients  $k$ ,  $k_0$ ,  $k_1$  of the Laurent expansion series of transformation (1) need to be known for the determination of the slope of the lift coefficient, the position of the pressure centre  $x_F$  and  $y_F$ , and the moment coefficient  $C_{mF}$  relative to the pressure centre. The relations among them are

$$\left. \begin{aligned} C_L^0 &= \left( \frac{dC_L}{d\alpha} \right)_{\alpha=0} = 8\pi k, \\ x_F &= k_{0R} - l^2 \cos(\alpha_0 + \mu), \\ y_F &= k_{0I} + l^2 \sin(\alpha_0 + \mu), \\ C_{mF} &= 4\pi l^2 k \sin(2\alpha_0 + \mu), \end{aligned} \right\} \quad (42)$$

where  $k_{0R} + ik_{0I} = k_0$ ,  $l^2 e^{-i\mu} = k_{1R} + ik_{1I} = k_1$ .

The expressions for  $k$  and  $k_{or}$  are derived previously (see equations (6), (15), (20b) and (33)). Here the expressions for  $k_{OI}$ ,  $k_{1R}$ ,  $k_{1I}$  are supplemented.

From equation (1),

$$k_0 = \frac{1}{2\pi i} \oint \frac{z}{\xi} d\xi = \frac{1}{2\pi} \int_0^{2\pi} zd\theta$$

Equating the imaginary parts on both sides.

$$k_{0z} = \frac{1}{2\pi} \int_0^{2\pi} yd\theta = \frac{1}{2\pi} \int_0^{2\pi} y \frac{d\theta}{d\phi} d\phi$$

Substituting equations (5), (16) and (21) into the above equation, the second order approximate solution of  $k_{OI}$  is found:

$$k_{0z} = \epsilon \cdot a_0 + \epsilon^2 \cdot f_0 \quad (43)$$

Similarly from equation (1),

$$k_1 = \frac{1}{2\pi i} \oint zd\xi$$

Equating the real parts and imaginary parts on both sides respectively,

$$\left. \begin{aligned} k_{1R} &= -\frac{1}{2\pi} \int_0^{2\pi} (x\cos\theta - y\sin\theta) d\theta, \\ k_{1I} &= -\frac{1}{2\pi} \int_0^{2\pi} (x\sin\theta + y\cos\theta) d\theta. \end{aligned} \right\} \quad (44)$$

For the two integral equations in (44), explicit expressions of second order accuracy could be obtained by means of the relations given previously. But the final forms are quite

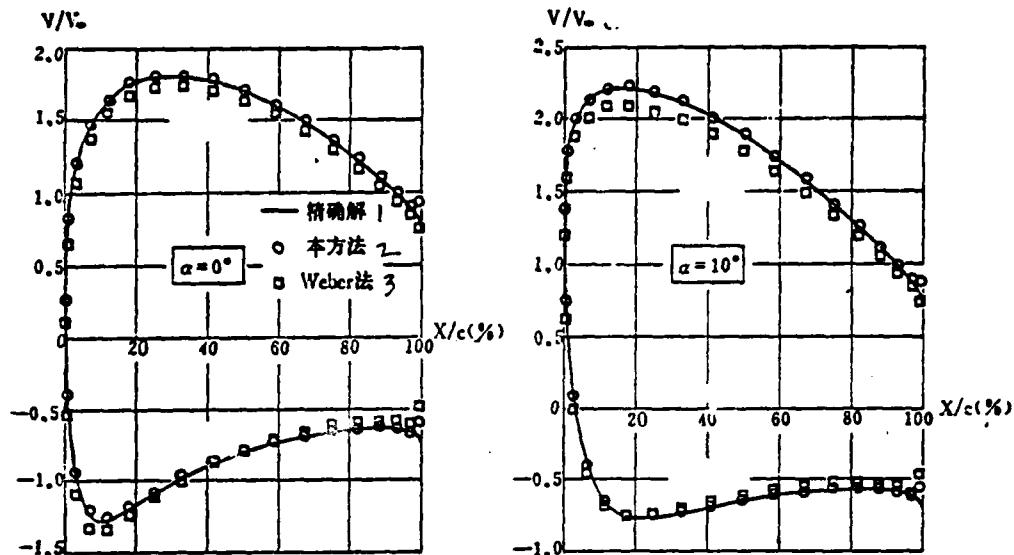
complicated. Hence direct approximation integration is recommended. The accuracy of the results is even better.

#### VIII. NUMERICAL EXAMPLES AND DISCUSSION

For evaluating of the accuracy of this method, calculations are executed for four different airfoils, and the results are compared with the exact solutions. The four airfoils are:

- (1) Piercy symmetric airfoil of 50% thickness ratio.  
Its exact solution is given by reference [7].
- (2) Zhukovskiy symmetric airfoil of 20% thickness ratio.
- (3) Zhukovskiy airfoil of 10% thickness ratio and 4% camber.
- (4) Zhukovskiy airfoil of 30% thickness ratio and 10% camber ratio.

The exact solutions of the three Zhukovskiy airfoils are also evaluated by the present author.



1 - calculated; 2 - exact solution; 3 - Weber's method.



The velocity distribution on the surface of the Zhukovskiy airfoil with 30% thickness ratio and 10% camber ratio.

The figure above shows the calculated results and the comparison with the exact solution of the fourth airfoil. From the figure, it is apparent that even for airfoils of such large thickness ratio and camber ratio, the accuracy of the second order approximation theory is still quite satisfactory. It is obviously even better than the Weber's method. For the first three airfoils, the accuracy of the calculated results from the second order approximation theory almost coincide perfectly with the exact solutions.

Finally, it is also noted that the compressibility modification can be easily introduced, as in reference [9], into

both the second order approximation theory presented by this paper and the formulae given by the improved Weber's method of reference [11] to calculate the velocity and the pressure distribution on an airfoil surface in a flow field of sub-critical free stream velocity. The author has been working on this problem and has already obtained some encouraging results. Hopefully the results will be presented in the near future for the readers' reference.

#### REFERENCES

- [1] Theodorsen, T., and Garrick, I. E., General Potential Theory of Arbitrary Wing Sections, NACA TR No. 452, 1933.
- [2] Glauert, H., The Elements of Aerofoil and Airscrew Theory, Cambridge University Press, 1926.
- [3] Goldstein, S., Approximate Two-dimensional Aerofoil Theory, Parts I-IV, ARC CP No. 128, 140, 141, 1952.
- [4] Lighthill, M. J., A New Approach to Thin Aerofoil Theory, Aero. Quart., Vol. 3, 1951, pp. 193-210.
- [5] Weber, J., The Calculation of the Pressure Distribution over the Surface of Two-dimensional and Swept Wings with Symmetrical Aerofoil Sections, ARC RM No. 2918, 1956.
- [6] Weber, J., The Calculation of the Pressure Distribution on the Surface of Thick Cambered Wings and the Design of Wing with Given Pressure Distribution, ARC RM No. 3026, 1957.
- [7] Spence, D. A., and Routledge, N. A., Velocity Calculations by Conformal Mapping for Two-dimensional Aerofoils, ARC CP No. 241, 1956.
- [8] Van Dyke, M. D., Second-order Subsonic Airfoil Theory Including Edge Effects, NACA TR No. 1274, 1956.
- [9] Lock, R. C., Wilby, P. G., and Powell, B. J., The Prediction of Aerofoil Pressure Distribution for Sub-critical Viscous Flows, Aero. Quart., Vol. 21, Part 3, 1970, pp. 291-302.
- [10] Milne Thomson, L. M., Theoretical Aerodynamics, London, 1952.
- [11] Lin, C., Some investigations and improvements on the Weber's method, unpublished, January, 1977.

## Aerodynamic Calculations and Design of Subcritical Aerofoils

*Lin Chaoqiang*

The present paper deals with both the direct problem (predicting pressure distribution for a given aerofoil at given angle of attack) and the inverse problem (finding aerofoil geometry and angle of attack for given pressure distribution) of aerofoils in subcritical potential flow. The emphasis is placed on the inverse problem, which is attacked perhaps somewhat more successfully than by existing methods.

The direct problem is considered first. Starting from the second order approximation theory previously obtained for calculating velocity distribution around an arbitrary aerofoil in incompressible potential flow<sup>[1]</sup>, the author suggests the compressibility correction, which is similar to that of reference [13] and predicts the aerodynamic character of aerofoils in subcritical potential flow. Compared with the solution of a specific aerofoil at critical Mach number, for which the exact solution exists, the accuracy of the formula is quite satisfactory and is better than that of Tsien's<sup>[11]</sup> and Lock's.<sup>[12]</sup>

The accuracy and fairly short computing time of the above obtained formula for the direct problem make it, when used in combination with the Newton iteration method, successfully applicable to the inverse problem, i.e., to the designing of an aerofoil with given pressure distribution along the chord at subcritical speed. The difference quotients are used to substitute for the partial derivatives which are needed in the iteration scheme. It is emphasized that these substitutions can be successfully used because the above mentioned formula can give not only accurate velocity distribution for direct problem itself, but also quite accurate difference of the velocity distributions of two aerofoils having sufficiently small difference in only one of the independent variables in the direct problem. As to convergence, it takes only 5~10 iterations to design an aerofoil with given pressure distribution.

Another interesting advantage of the present aerofoil design technique is that it can also be used, without additional difficulty, in the so-called mixed design problems, such as the design of an aerofoil with given thickness distribution and upper surface pressure distribution.

## I. INTRODUCTION

Based on the previous works by the author, references [1] and [2], this paper proposes a method of calculating the velocity and pressure distribution on an airfoil of given geometrical profile at a given angle of attack in subcritical potential flow and, inversely, designing appropriate airfoil geometry and the corresponding angle of attack for a given pressure distribution.

There already exist a number of methods for the calculation of the pressure distribution on an airfoil in incompressible flow, for instance, references [3]-[10]. The author proposed an approximation method in reference [1] which is relatively simple. On the other hand, it produces rather satisfactory results within practical ranges of thickness ratio and camber ratio of common airfoils. It practically coincides with the exact solution. Based on the second order approximation theory of reference [1], reference [2] further shows that, if only the first order term is maintained and the higher order terms are neglected, then results similar to those given by Weber's method [7], [8] are obtained. Hence, this method can be referred to as the improved Weber's method or, from the above point of view, the first order approximation theory. The numerical examples in reference [2] indicate that it provides the same degree of accuracy as the Weber method, but is slightly improved compared with the Weber method. Hence, we suggest that for all calculations proposed by the Weber method, the method given by reference [2] can be employed as a substitute.

For the compressibility correction, there exists the early well known Kerman-Tsien correction formula [11], [12]. Discussions by many papers appeared later, for instance, references [13]-[15]. The results given by reference [13] indicated that using Equation (4c) of that paper as a compressibility correction, the pressure distribution of the airfoil thus obtained agrees well with the exact solution, except in the vicinity of the leading edge where the error is quite significant. In reference [13], the Weber method

was employed for the calculation of the corresponding incompressible flow. The paper suggests that the significant error near the leading edge is due to the fact that the Weber method is inaccurate near the leading edge of the airfoil under the condition of incompressible flow. In this paper, we will apply the second order approximation [1] instead of Weber's method in reference [13], while the method for compressibility correction given by reference [13] is adopted. The results obtained from these calculation procedures show that the accuracy is greatly improved, especially in the vicinity of the leading edge.

For the design of airfoils, Lighthill [16] publicized an important paper proposing a very good design method in 1945. Its fundamental concept was employed by many authors afterwards, for instance, references [17]-[19]. The method of airfoil design presented here is very simple and direct in principle. Based on the solution of the direct problem (that is, predicting velocity and pressure distribution on an airfoil surface for a given profile and an angle of attack), the solution can be sought by means of the Newton iteration method. In the Newton method, the partial differentials are approximated by the quotient of the corresponding finite differences. The calculation time required to design an airfoil with such an iteration method solving a set of simultaneous non-linear equations is much longer for calculating a direct problem, since each iteration is equivalent to calculating tens of direct problems (depending upon the number of discrete points of the airfoil coordinates). However, this disadvantage does not cause any difficulties at all because the solution for the direct problem obtained previously is very accurate and requires only a very short calculation time. Generally, only 5-10 iterations are sufficient to obtain convergent solutions. Another important characteristic of the solution of the direct problem is: besides being accurate in itself, the corresponding difference in surface velocity due to a slight difference of any independent variable of the direct problem is also accurate. Hence, it is reasonable to

approximate partial differentials by the quotients of finite differences.

The other advantage of this method is that it can be applied to the mixed design problems of the airfoil conveniently--that is, design problems for which the geometrical profile is given partially and the velocity profile is given partially. For example, one design problem is obtaining the camber distribution of the airfoil and the corresponding angle of attack given the thickness distribution of the airfoil and the surface velocity distribution. Such a problem is of practical use since a certain thickness should be prescribed to satisfy strength and structural requirements, while, on the other hand, the main purpose of airfoil design in general is to ensure a certain velocity on the upper surface.

In the second section of this paper, the compressibility correction based on the methods given by references [1] and [2] (that is, the calculation method for velocity distribution on the upper surface in subcritical flow) is discussed. In the third section, the airfoil design procedure by means of the Newton method for a given surface velocity or pressure distribution will be discussed. The fourth section deals with the mixed design problem, that is, the method of designing an airfoil for a given thickness distribution and a prescribed velocity distribution on the upper surface. The fifth section is a conclusion.

## II. AERODYNAMIC CALCULATIONS OF SUBCRITICAL AIRFOILS

Here the method of calculating the velocity distribution on the surface of an airfoil in subcritical flow for a given geometrical profile and an angle of attack is discussed.

First the formula for calculating the velocity distribution on the surface of an airfoil in incompressible flow given by references [1] and [2] should be introduced.

The coordinate system is taken as shown in Figure 1. The chord length of the airfoil is set at unity. The leading edge falls on the origin and the trailing edge is found at the point (1,0).

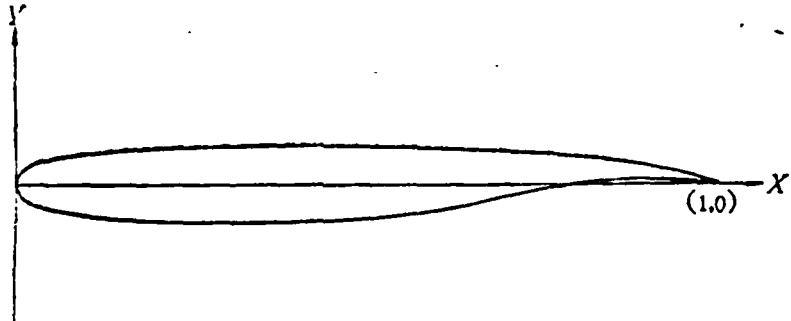


Figure 1. Coordinate system of the airfoil

Let  $x = \frac{1}{2}(1 + \cos \phi)$ , (2.1)

in which the upper surface corresponds to a variation of  $\phi$  from 0 to  $\pi$ , while the lower surface corresponds to a variation of  $\phi$  from  $\pi$  to  $2\pi$ . The profile of the airfoil is assumed to be given by the  $(x, y)$  coordinates of  $2N$  discrete points on the airfoil surface, that is, by taking the following values of the  $x$  coordinate:

$$x_j = \frac{1}{2}(1 + \cos \phi_j), \quad \phi_j = j\pi/N, \quad (j = 0, 1, 2, \dots, 2N-1, 2N) \quad (2.2)$$

and then the corresponding values of the  $y_j$  coordinates. For the numerical examples in this paper, we take  $N = 13$ .

It is now assumed that the profile of the airfoil can be approximated by the following trigonometric polynomial:

$$y = \epsilon \left\{ a_0 + \sum_{n=1}^{N-1} (a_n \cos n\phi + b_n \sin n\phi) + a_N \cos N\phi \right\}, \quad (2.3)$$

where  $\epsilon$  is a small quantity of the same order of magnitude as the thickness ratio and the camber ratio, and the coefficients  $a_n$  and  $b_n$  can be determined by equating the  $y$  values of the discrete points defined by equation (2.3) and the given values.

The formula for calculating the velocity distribution on an airfoil surface in incompressible flow given by reference [1] is (setting the free-stream velocity equal to unity)

$$V_i = \frac{4k \sin \frac{\theta - \alpha_0}{2} \cos \left( \alpha - \frac{\theta + \alpha_0}{2} \right) \cdot \left| \frac{d\xi}{d\phi} \right| / \left| \frac{d\xi}{d\theta} \right|}{\sqrt{\left( \frac{dx}{d\phi} \right)^2 + \left( \frac{dy}{d\phi} \right)^2}}, \quad (2.4)$$

where  $\theta$  is the angular position of the point on the unit circle corresponding to a point on the airfoil after transformation,  $\alpha$  is the angle of attack,  $\alpha_0$  is the constant zero-lift angle of attack, and the meanings of  $k$  and  $\xi$  are defined in reference [1]. All the quantities in the above expression can be computed one by one with the values of the coordinates at the given discrete points and the formulae given in reference [1].

If only the first order of  $\epsilon$  is preserved in Equation (2.4), the following result is obtained according to the derivation given by reference [2]. It is similar to the results obtained by Weber's method, and hence it is also called the first order approximation method or the improved Weber method:

$$V_i = \frac{\cos \alpha \{ \pm \sin \phi \pm T^{(1)}(x) + C^{(1)}(x) \} + \sin \alpha \{ (1 - \cos \phi) + T^{(1)}(x) \pm C^{(1)}(x) \}}{\sqrt{\sin^2 \phi + [T^{(1)}(x) \pm C^{(1)}(x)]^2}}, \quad (2.5)$$

in which  $T^{(1)}(x)$ ,  $T^{(2)}(x)$ ,  $T^{(3)}(x)$ ,  $C^{(1)}(x)$ ,  $C^{(2)}(x)$  and  $C^{(3)}(x)$  are all linear functions of the thickness ratio and the camber ratio of the airfoil.

Now the compressibility correction is made according to Equation (4c) of reference [13].

When the velocity distribution on the surface of the airfoil in incompressible flow is calculated according to Equation (2.4), the following expression is employed for the compressibility correction:

$$V = \frac{\frac{1}{2} \sin \phi + \frac{1}{B} \left\{ 4k \sin \frac{\theta - \alpha_0}{2} \cos \left( \alpha - \frac{\theta + \alpha_0}{2} \right) \left| \frac{d\xi}{d\phi} \right| / \left| \frac{d\xi}{d\theta} \right| - \frac{1}{2} \sin \phi \right\}}{\sqrt{\left( \frac{dx}{d\phi} \right)^2 + \frac{1}{B^2} \left( \frac{dy}{d\phi} \right)^2}}, \quad (2.6)$$

where

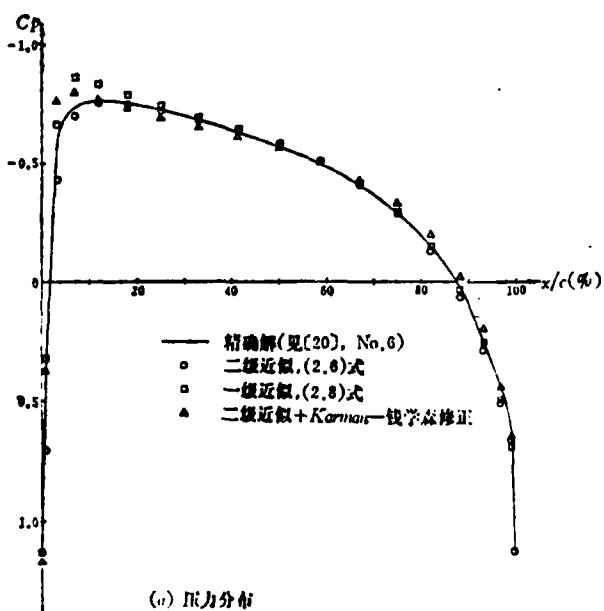
$$B = \sqrt{1 - M_\infty^2(1 - M_\infty C_{p1})},$$

$M_\infty$  is the free stream Mach number, and  $C_{p1}$  is the local pressure

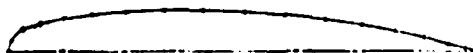
coefficient of the incompressible flow.

When Equation (2.5) is used to calculate the velocity distribution on the airfoil in incompressible flow, the following expression is applied for compressibility correction:

$$V = \frac{\cos \alpha \left\{ \pm \sin \phi + \frac{1}{B} \left[ \pm T^{(1)}(x) + C^{(1)}(x) \right] \right\} + \frac{\sin \alpha}{B} \left\{ (1 - \cos \phi) + \sqrt{\sin^2 \phi + \frac{1}{B^2} \left[ T^{(2)}(x) \pm C^{(2)}(x) \right]^2} \right.}{\left. + \frac{1}{B} \left[ T^{(3)}(x) \pm C^{(3)}(x) \right] \right\} + \frac{1}{2} \frac{a^2}{B^2} \sin \phi}$$



(a) 压力分布  
(b) Profile of the airfoil



• 按二级近似理论设计  
(b) 翼型几何形状

Figure 2. The pressure distribution and the profile design of NLR subcritical airfoil,  $M_\infty = 0.7032$ ,  $\alpha = 0^\circ$

- exact solution (see reference [20], no. 6)
- second order approximation, Equation (2.6)
- first order approximation, Equation (2.8)
- △ second order approximation + Karman-Tsien correction
- design according to the second order approximation theory

### III. DESIGN OF THE AIRFOIL

In this section, the design of the geometrical profile of the airfoil and the angle of attack for given velocity distribution and pressure distribution on the airfoil surface is discussed.

It is assumed that the values of the velocity on the upper and the lower surface are given at the  $2N-1$  points ( $j = 1, 2, \dots, 2N-1$ ) according to Equation (2.2), and the Newton iteration method is used to determine the thickness ratio of the airfoil, the camber ratio of the airfoil and the angle of attack. The procedure is discussed briefly as below.

The problem consists of  $2N-1$  unknowns, that is,  $N-1$  thickness ratio values,  $N-1$  camber ratio values and the angle of attack  $\alpha$ . Let  $y_1, y_2, \dots, y_{2N-1}$  represent these  $2N-1$  unknowns. Then the velocity on the airfoil surface  $U_j$  can be expressed as functions of  $y_1, y_2, \dots, y_{2N-1}$ , that is

$$U_j = f_j(y_1, y_2, \dots, y_{2N-1}), \quad (j=1, 2, \dots, 2N-1) \quad (3.1)$$

The selection of the initial values of the iteration can be quite arbitrary, for example,  $\alpha = 0^\circ$  and an ellipse of 10% thickness ratio can be taken. The iteration must finally satisfy

$$|U_j - V_j| < \epsilon, \quad (3.2)$$

where  $\epsilon$  is the control error of the iteration. For the numerical examples in this paper,  $\epsilon = 10^{-4}$  is assumed.

Assuming that the unknown of the  $k$ th iteration is  $y_i^{(k)}$ , ( $i = 1, 2, \dots, 2N-1$ ), and by letting

$$y_i^{(k+1)} = y_i^{(k)} + \Delta y_i^{(k)}, \quad (i=1, 2, \dots, 2N-1) \quad (3.3)$$

then the value of  $\Delta y_i^{(k)}$  is determined by the following system of linear algebraic equations according to the Newton method:

$$U_j^{(k)} + \sum_{i=1}^{2N-1} \frac{\partial f_j}{\partial y_i} \Delta y_i^{(k)} = V_j, \quad (j=1, 2, \dots, 2N-1) \quad (3.4)$$

The values of the partial derivatives in the above equation are calculated approximately by the finite difference forms:

$$\frac{\partial f_i}{\partial y_i} \cong \frac{1}{h} \left\{ f_i(y_1, y_2, \dots, y_{i-1}, y_i + h, y_{i+1}, \dots, y_{2N-1}) - f_i(y_1, y_2, \dots, y_{i-1}, y_i, y_{i+1}, \dots, y_{2N-1}) \right\}, \quad (3.5)$$

For the numerical examples presented in this paper,  $h = 0.5 \times 10^{-4}$  is taken. Generally, only 5-10 iterations are required to achieve convergence.

Figure 2(b) shows the design result for the sixth airfoil shape of reference [20] by the second order approximation theory presented in this paper. The free stream Mach number is  $M_\infty = 0.7032$ . The pressure coefficient at the point of minimum pressure is just equal to the critical pressure coefficient. It can be observed from the figure that, for the condition of symmetrical subcritical airfoil, the airfoil profile designed by the second order approximation theory is very accurate. Even the design result by the first order approximation theory (not shown in the figure) is still quite satisfactory, although it is less accurate than that produced by the second order approximation theory.

Figure 3 is an example of a low drag-high lift airfoil adopted from Figure 10 of reference [17]. The pressure coefficient distribution in Figure 3 is sketched according to Figure 10 of [17]. Hence, there might be a slight deviation from the original distribution. The airfoil profiles in Figure 3 are designed by the second order and first order approximation theories, accordingly. Figure 10 of reference [17] indicates that the thickness ratio of the airfoil is 18%, and the angle of attack is  $\alpha = 11.2^\circ$ . According to the result computed by the second order approximation theory, the thickness ratio is 18.5% and the angle of attack is  $\alpha = 11.5^\circ$ . Furthermore, it can be observed directly that the second order approximation airfoil profile given in Figure 3 resembles very closely the one shown in Figure 10 of reference [17]. Considering

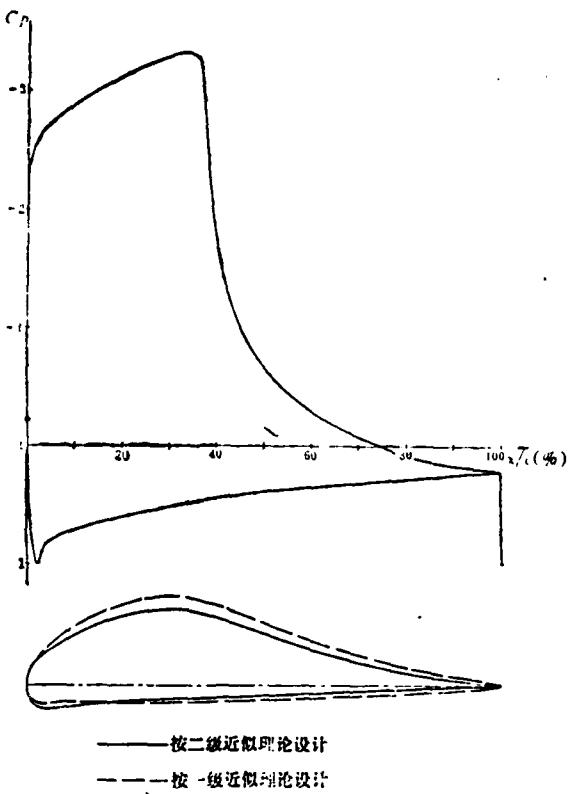


Figure 3. Design of low drag-high lift airfoil  
 ——design according to the second order approximation theory  
 -----design according to the first order approximation theory

that there is a certain unavoidable error of the data for the pressure coefficient distribution, we assume that the result given by the second order approximation theory is more satisfactory. The first order approximation theory predicts a thickness ratio of 22% and an angle of attack of  $\alpha = 11.8^\circ$  and has a greater error. Similar to the airfoil computation, problem 1, there is a greater error in the results provided by the first order approximation theory for airfoils with relatively large camber.

#### IV. THE MIXED DESIGN PROBLEM FOR THE AIRFOIL

The so-called mixed design problem is characterized by partially specifying the geometrical shape of the airfoil and the velocity distribution on the airfoil surface, and then determining

the remaining geometrical profile, velocity distribution and the angle of attack. For example, assuming that the thickness ratio of the airfoil and the velocity distribution on the upper surface are given, the camber distribution of the airfoil and the angle of attack is to be determined. This problem can be treated almost in the same way as proposed in the previous section. The only difference is that the number of unknowns is fewer (by  $N-1$  thickness values). Figure 4 is the design result according to the exact solution data of a 10% thickness ratio, 4% camber ratio Zhukovskiy airfoil at  $M_\infty = 0$ .

For the exact solution, the camber ratio of the airfoil is given as 4%, and the angle of attack is  $\alpha = 10^\circ$ . According to the result from second order approximation computations, the camber ratio is 3.98% and the angle of attack is  $\alpha = 9.97^\circ$ , which are very close to the exact solution. From the result of the first order approximation computation, the camber ratio is found to be 4.41%, and the angle of attack is  $\alpha = 10.24^\circ$ . The solid line in Figure 4 shows the exact airfoil profile, while the dots are the results of the second order approximation. The results of the first order approximation are not shown.

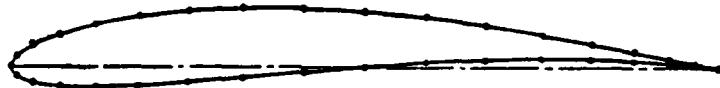


Figure 4. Mixed design of the airfoil for given thickness ratio and given pressure distribution on the upper surface.

## V. CONCLUSION

Based on the second order approximation theory of reference [1] and the first order approximation theory of reference [2], and adopting the compressibility correction method of reference [13],

this paper has proposed a method of computing the velocity distribution on an airfoil in subcritical flow according to the second order approximation theory of Equation (2.6) or the first order approximation theory of Equation (2.8). Comparison with the results from the exact solution shows that these two methods, especially the second order approximation Equation (2.6), are very accurate and reliable for subcritical Mach numbers.

Based on the Equations (2.6) or (2.8), this paper proposes using the Newton iteration method to design the profile of the airfoil and to find the corresponding angle of attack for given chordwise velocity distribution on the airfoil surface. For airfoils with small camber, both the first order approximation and the second order approximation design methods are fairly accurate. For airfoils with relatively larger camber, the error of the first order approximation method is relatively greater. However, the second order approximation method still has satisfactory accuracy.

The methods presented in this paper can also be applied to the mixed design problem of the airfoil, for example, the problem of airfoil design for given thickness distribution of the airfoil and given velocity distribution on the upper surface.

## REFERENCES

- [1] Lin, Chaoqiang, Second Order Approximation Theory of an Arbitrary Airfoil in Incompressible Potential Flow, Transactions of Northwest Technical University, 1979, pp 15-25.
- [2] Lin, Chaoqiang, Some Investigations and Improvements on the Weber Method, unpublished, January 1977
  - [3] Theodorsen, T., and Garrick, I. E., General Potential Theory of Arbitrary Wing Sections, NACA TR No. 452, 1933.
  - [4] Glauert, H., The Elements of Aerofoil and Airscrew Theory, Cambridge University Press, 1926.
  - [5] Goldstein, S., Approximate Two-dimensional Aerofoil Theory, Parts I-IV, ARC CP No. 128, 140, 141, 1952.
  - [6] Lighthill, M. J., A New Approach to Thin Aerofoil Theory, Aero. Quart., Vol. 3, 1951, pp. 193-210.
  - [7] Weber, J., The Calculation of the Pressure Distribution over the Surface of Two-dimensional and Swept Wings with Symmetrical Aerofoil Sections, ARC RM No. 2918, 1956.
  - [8] Weber, J., The Calculation of the Pressure Distribution on the Surface of Thick Cambered Wings and the Design of Wing with Given Pressure Distribution, ARC RM No. 3026, 1957.
  - [9] Spence, D. A., and Routledge, N. A., Velocity Calculations by Conformal Mapping for Two-dimensional Aerofoils, ARC CP No. 241, 1956.
  - [10] Van Dyke, M. D., Second-order Subsonic Airfoil Theory Including Edge Effects, NACA TR No. 1274, 1956.
  - [11] Tsien, H. S., Two-dimensional Subsonic Flow of Compressible Fluids, JAS Vol. 6, No. 10, 1939, pp. 339-407.
  - [12] Von Karman, Th., Compressibility Effects in Aerodynamics, JAS Vol. 8, No. 9, 1941, pp. 337-356.
  - [13] Lock, R. C., Wilby, P. G., and Powell, B. J., The Prediction of Aerofoil Pressure Distribution for Sub-critical Viscous Flows, Aero. Quart., Vol. 21, Part 3, 1970, pp. 291-302.
  - [14] Labrujere, T. E., Loeve, W., and Slooff, J. W., An Approximate Method for the Determination of the Pressure Distribution on Wings in the Lower Critical Speed Range, AGARD CP. 35-17, 1968.
  - [15] Van Dyke, M. D., The Second-order Compressibility Rule for Airfoils, JAS Vol. 21, No. 9, 1954, p. 647.
  - [16] Lighthill, M. J., A New Method of Two-dimensional Aerodynamic Design, ARC RM No. 2112, 1945.
  - [17] Liebeck, R. H., A Class of Airfoils Designed for High Lift in Incompressible Flow, J. Aircraft Vol. 10, No. 10, 1973, pp. 610-617.
  - [18] Strand, T., Exact Method of Designing Airfoils with Given Velocity

Distribution in Incompressible Flow, J. Aircraft Vol. 10, No. 11, 1973, pp. 651-659.

[19] Nonweiler, T. R. F., A New Series of Low-drag aerofoils, ARC RM No. 3618, 1971.

[20] Boerstoel, J. W., Symmetrical Subsonic Potential Flows Around Quasi-elliptical Aerofoil Sections, NLR TR 68016 U, 1968.

1978年1月

# An Aerodynamic Design Method for Transonic Axial Flow Compressor Stage

*Zhu Fangyuan, Zhou Xinhai,  
Liu Songling, and Fan Feida*

A three dimensional aerodynamic design method for transonic axial flow compressor stage is described in detail in this paper in order to make it easier to apply and more widely used. The method comprises three main parts: the mean  $S_1$  streamsurface calculation, the approximate calculation of  $S$  streamsurface of revolution, and defining the blade element on the conical surface and stacking the blade airfoil sections. The method is unusual in that the calculation stations for making the  $S_1$  streamsurface computations are curves, and particularly in that the airfoil parameters of blade are calculated on a plane tangent to the approximate streamsurface of revolution. On this tangential plane, two dimensional flow is used as a basic model to calculate the Mach wave system on the suction surface of cascade entrance region.

The streamline curvature method is used to calculate the flow field on mean  $S_1$  streamsurface. The projections on meridional planes, of the blade leading and trailing edges, are selected as calculation stations. Along curved calculation stations, the principal equations, in which the streamline curvature and the gradients of enthalpy and entropy are taken into account, are derived from the fundamental equations of non-viscous axisymmetric flow. The Runge-Kutta method is used to solve the principal equations. The slope and curvature of the streamline are found by means of the spline and double spline functions respectively.

The approximate calculation of  $S_1$  streamsurface of revolution consists of the free stream calculation and the blade airfoil parameters calculation. The free stream in cascade entrance region is calculated for the purpose of performing the calculation of unique incidence angle and the analysis of choking margin of blade channel. In the free stream calculation, the continuity equation is used to obtain the flow parameters, and the basic assumption adopted is that the entropy and  $V_{\infty}r$  are constant on each streamline.

The multiple-circular-arc (MCA) airfoils are used for both the rotor and the stator. The parameters of MCA airfoils are calculated on a plane which

is tangent to the approximate conical streamsurface. On the tangential plane the airfoil section is defined by specifying the following parameters: the blade setting angle, the length of the front chord, the length of the total chord, the front camber, the total camber, the maximum thickness and its location, and the radii of the leading and the trailing edge. These parameters should be so adjusted as to be compatible with the required values of the incidence angle, of the blade channel choking margin, and of the distance from the leading edge to the assumed normal shock impinging point on the suction surface.

In the third part of the present method, the previously obtained MCA airfoils are transferred from the tangential plane to the developed conical streamsurface, on which the circular arc is approximated by the constant turning rate curve. The calculation methods used in this step and used to stack the designed airfoil sections are the same as those in reference[8].

The calculated results of the present method are compared with those of references[2][3] and found in satisfactory agreement. Therefore, the present method appears to be a useful tool for the aerodynamic design of transonic axial flow compressor stage.

# AN AERODYNAMIC DESIGN METHOD FOR TRANSONIC AXIAL FLOW COMPRESSOR STAGES

Zhu Fanyuan, Zhou Xinhai, Liu Songling and Fan Feida

## 1. INTRODUCTION

At the end of the 1950's, Professor Wu Zhong Jiao put forward a theory of the aerodynamics of three dimensional flow in a turbine wheel mechanism and by going through successive solutions for  $S_1$  and  $S_2$  stream surfaces, he was able to solve for the three dimensional flow field; this laid the theoretical foundation for the use of three dimensional aerodynamic models in the aerodynamic design of turbine machinery.

By the end of the 1960's, there was a continuous stream of reports published outside China--these included reports [1]-[7]-- which introduced the use of three-dimensional theory for the design of transonic compressor stages and the results of calculations and experiments using this procedure. These transonic stages all made use of multiple-circular-arc airfoils (MCA). According to the data which was published in these reports, it was possible to see that there was a relatively good agreement between the experimental values and the design values for the average capabilities as well as the distribution of aerodynamic parameters along the radial direction. This demonstrated that, in the designing of transonic compressor stages, the use of the three-dimensional theory was capable of achieving excellent results. This also demonstrated that the multiple circular arc airfoil is one type of airfoil which is available for use in the designing of transonic stages.

On the basis of analytical research in the references stated above, we created an actual aerodynamic design method for use in the case of transonic axial flow compressor stages; moreover, we also wrote a computer program based on this method.

This computational method includes:

- (1) Average  $S_2$  stream surface calculation--blade leading edge and trailing edge flow field calculation.

If we make use of the flow line rate of curvature method, place the calculation stations on the leading and trailing edges of blades, and take the projection of the calculation stations on the meridional planes, these projections define a certain curve. In these calculations, we have taken into account the influences of enthalpy, entropy and the rate of curvature of the flow lines. The result of these calculations are the aerodynamic parameters for the average  $S_2$  stream surface on the leading and trailing edges of blades involved in the measurements.

- (2) The approximate calculation of the stream surface of revolution  $S_1$ .

At present, the actual procedure which is used in design is to make use of, under selected conditions of airfoil shape, an approximate calculation of the  $S_1$  stream surface; the purpose of this is to make use of the approximate calculation of  $S_1$  in order to check out the aerodynamic characteristics of the airfoil type selected. When one employs this method, the approximate calculation of the  $S_1$  stream surface includes:

#### A. Free stream calculation

The free stream calculations are a component part of the approximate calculation of the cascade entrance flow field of the  $S_1$  stream surface. What this means is that one takes a reference position at which the  $S_1$  stream surface and the leading edge aerodynamic parameters are already known, and then, under conditions in which there is equal entropy and the values for  $C_s$  do not vary, one calculates, from positions of the

leading edge, the distribution of parameters along the flow lines. This is done so that when one is calculating the airfoil parameters, it will be possible to check out the original data for unique incidence angles and the choke margins of the blade channels. In such calculations, the effects of changes in position of flow lines in a radial direction as well as changes in the stream surface thickness as they impact on aerodynamic parameters are taken into account.

#### B. Calculation of parameters of multiple arc airfoils.

In this type of calculation, one makes use of an approximately conic surface in order to replace the surface of rotation of  $S_1$ . Then one calculates the multiple arc airfoil parameters on a plane which intersects this conic surface. One must distinguish how they relate to angle of incidence (unique angle of incidence or leading edge angles of incidence), the locations of points of intersection between the trough shock wave and the back edge of the airfoils and the choking margin of the blade cascade passage-way; these quantities are requirements for the precise determination of the angle of incidence of the center line of the multiple arc airfoil, the trailing angle, the ratio between the front chord length and the total length, and the ratio between the front camber and the total camber. If we add some other given conditions, then it is possible to make precise determinations for a multiple arc airfoil. In these calculations, changes along the flow lines of the blockage parameters which correspond to various flows as well as changes in the thickness of the stream surface were considered.

#### (3) Blade configuration and successive integration

On the basis of the multiple arc airfoil parameters which have already been obtained through calculations, the blade type under consideration can be configured on a developed conical surface. In order to be able to maintain the characteristics of

the multiple arc body as a plane figure on the conical surface we have mentioned, use can be made of a constant turning rate curve in order to simulate the multiple arcs. After configurations have been completed for the various high blade stream surfaces, then on the basis of definite requirements, one can take the various integrations of the airfoil form and use them to form the blade in question; moreover, in the same fashion, one can obtain the coordinates for the surface shapes of the various cross-sections of the construction. The principles behind this dividing operation as well as the computational sequence involved are both completely the same as [9].

## 2. THE CALCULATION OF THE FLOW FIELDS OF THE LEADING AND TRAILING EDGES OF BLADES

On the basis of the continuity equations for non-viscous fluids as well as their momentum equations and state equations, if one works according to the related equations (Figure 1):

$$\frac{\partial}{\partial l} = \cos \lambda \frac{\partial}{\partial r} + \sin \lambda \frac{\partial}{\partial z} \text{ and } \frac{\partial}{\partial m} = \sin \varphi \frac{\partial}{\partial r} + \cos \varphi \frac{\partial}{\partial z},$$

then it is possible to deduce the primary equation for the computational station, l, on a curve by using the flow line rate of curvature method, that is,

$$\begin{aligned} \frac{\partial C_m}{\partial l} = & \frac{gR}{C_m} \left( -\frac{K}{K-1} \frac{\partial T^*}{\partial l} + \frac{T^*}{\sigma} \frac{\partial \sigma}{\partial l} \right) - \frac{C_s}{C_m} \frac{\partial (rC_s)}{\partial l} - \frac{C_s^2}{C_m} \frac{K-1}{2K} \frac{1}{\sigma} \frac{\partial \sigma}{\partial l} \\ & - \frac{C_m}{1-M^2} \left[ \frac{1+M^2}{r} \sin \varphi \sin(\varphi + \lambda) + \frac{\sin(\varphi + \lambda)}{R_m} \operatorname{tg}(\varphi + \lambda) \right. \\ & \left. + \operatorname{tg}(\varphi + \lambda) \frac{\partial \varphi}{\partial l} \right] - C_m \frac{\cos(\varphi + \lambda)}{R_m} - C_m \frac{K-1}{2K} \frac{1}{\sigma} \frac{\partial \sigma}{\partial l}, \end{aligned} \quad (2.1)$$

$$\frac{\partial G_s}{\partial l} = 2\pi gr\rho C_m \cos(\varphi + \lambda), \quad (2.2)$$

In this equation  $\sigma = e^{-(\theta-\theta_0)/4B}$ . A is the amount of heat which corresponds to the power involved.

In the computational program involved here, use is made of the Runge-Kutta method for solving the set of equations (2-1) and (2-2); by this solution it is possible to obtain the meridional speed,  $C_m$ , along the calculation stations as well as the distribution of amounts of flow between the computational points and the base section.

If we first make use of the double spline function to deal with the projection of flow lines on the meridional plane and we then make initial use of the single spline function to deal with the points representing equal amounts of flow, then one can solve for the rate of slope of the flow lines involved,  $\operatorname{tg}\psi$ , and make a second use of the single spline function in order to deal with the rate of slope of these flow lines and solve for the rate of curvature of the flow lines as well.

Concerning the estimation of losses due to the blades, in the computational sequence which we are using there are two types of methods which offer a way for selecting good blade arrangements. One method is, on the basis of data which is already on hand, to determine the radial distributions for stator pressure recovery coefficients and rates of equal entropy for rotors; the other method is, on the basis of the relationship between given diffusion factors and loss coefficients, to estimate the airfoil losses and, at this time, to resubstitute into the calculations aerodynamic coefficients which have already been obtained; as far as the diffusion factors and loss coefficients which are substituted into these calculations go, when the flow fields converge, then the aerodynamic parameters and loss coefficients should also be the same (within a range of permissible deviation).

Tables 1 and 2 show the results of these calculations; moreover, [2] and [3] carry out a comparison of the data given.

### 3. FREE FLOW CALCULATIONS

The problem which must be solved when one is doing free flow calculations is this. When one is calculating airfoil parameters, one must find the neutral point on the back of the blades involved which is related to the unique angle of incidence to the blade cascade as well as having to calculate shock wave losses in order to supply basic data.

What is meant by the term free flow is this. From a position on the leading edge of a blade, when there is no effect from blades and there is only the influence of the shortening of the flow route in the meridional plane, then the only flow is due to this shortening in the meridional plane. Because of this fact, free flows should be able to satisfy the following conditions:

- (1) the  $C_{ar}$  values for free flows should be the same as those for aerodynamic flows at places on the leading edges of blades when those places are on the same flow lines as the free flows;
- (2) there should be equal entropies of flow and
- (3) the flows should be axisymmetrical.

When one is dealing with the existence of blades, in the area between the entrance to the cascade and the first covered surface of the trough (the neutral point on the back of the blades involved as well as the trough shock waves are all located in the intake area), due to the mutually offsetting functions of the shock waves and the expansion waves on leading edges and due the fact that before gas flow has entered the cascade troughs the

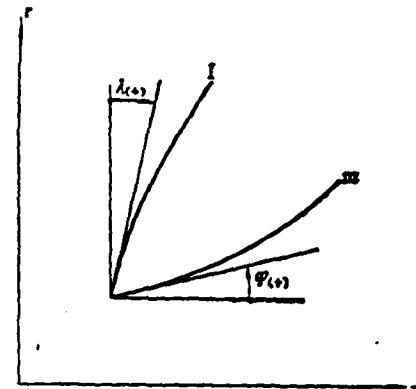


Figure 1

control which is exerted by the blades is relatively small, and also due to the fact that when one is dealing with supersonic speeds of oncoming flow the forward section of multiple arc airfoils is relatively flat, it is still possible to say that the average circumferential parameters of airflow closely satisfy the three conditions stated above.

Calculations of free flows are carried out on the foundation of average  $S_2$  stream surface calculations along flow lines or streamlines according to the conditions set out above and by the use of continuity equations.

In this calculation process, the important formulas are:

$$C_m = \frac{1}{2\pi g} \frac{1}{r\rho} \frac{1}{\cos(\varphi+\lambda)} \frac{dG_B}{dl}, \quad (3.1)$$

$$g\rho = \frac{p^*}{RT^*} \left(1 - \frac{K-1}{2} \frac{C_m^2 + C_\infty^2}{KgRT^*}\right)^{\frac{1}{K-1}}. \quad (3.2)$$

From these calculations, we can obtain values for  $M_\infty$ ,  $\beta_m$ ,  $p_m^*$ ,  $T_m^*$ ,  $\frac{\partial G_B}{\partial r}$ , and these values are useful when calculating airfoil parameters.

#### 4. THE CALCULATION OF MULTIPLE ARC AIRFOIL PARAMETERS

Using these types of design methods, the configuration for blades is worked out on a developed conical plane. The function of airfoil parameter calculations is to supply needed data for the configuration and successive integration of blades.

Let us now take several of the important questions which come up when one is doing airfoil parameter calculations and discuss them below.

(1) On what kind of plane should one calculate airfoil parameters?

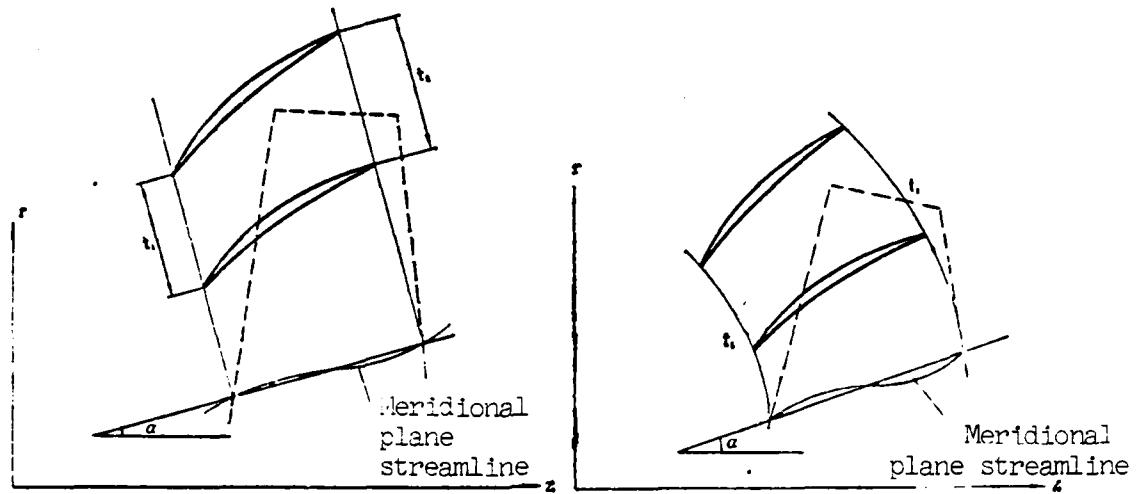


Figure 2

We recognize the fact that there are two workable methods of making this decision. One method is to make the decision using a tangential plane to a conic surface. The other method is to make the decision using a developed conic surface. The first of these methods is composed of two sections of arcs, one for the dish-shaped face of the blade and one for the suction surface or the back of the blade. The second of the two methods is made up of two sections of curve with equal rates of rotation, one for the dish-shaped side of the blade involved and one for the suction side on the back (see Figure 2).

When the oncoming flow corresponds to an  $M$  number in excess of 1.0, then when one is working on the basis of the principle of a unique angle of incidence, it is necessary to do calculations of the Mach system in the intake area of the cascade. When one does these calculations on a tangent plane to the conic surface, it is necessary to obtain the calculations for this Mach system by the use of a method which makes use of a two-dimensional flow model and then makes adjustment to compensate for the introduction of three dimensions.

If one is not calculating the design of the airfoil on the basis of the principle of a unique angle of incidence, but is making these calculations on the basis of the idea that the angles of incidence for the leading edges of the suction faces of the blades involved are chosen at a certain size, then it is not necessary to calculate the Mach system for the intake of the cascade. It is also possible to make use of the second method mentioned above and (9) is done in this way.

The computational sequence which we have been talking about makes use of the first method. Moreover, when one is making a check of the situation which exists when the cascade is blocked, then one considers changes in the cascade distance along the conical plane, to changes in the blockage parameters which correspond to different airflows, and to changes in the thickness of the stream surfaces involved. These calculations demonstrate that the various airfoil parameters which are obtained by this sort of procedure are relatively reasonable. The results of these calculations and the data from [3] are both set out in Table 3.

## (2) The calculation of the lag angle

Theoretically speaking, the calculations for lag angles are all based on the Kutta Formula and then adjustments are made to take the factor of a three-dimensional flow field into consideration. Generally speaking, there are two types of methods which can be used. One method is to make use of the flapping angle in order to do calculations such as (2) and (3). The other method is to make use of the equivalent flapping angle to make the calculations as (5) and (9).

### A. Calculations made by the use of the airfoil flapping angle.

The formula for the calculations is:

$$\begin{aligned}
 \delta &= \frac{m\phi}{\sqrt{\sigma}} + X \\
 &= \frac{\beta_{2m} - \beta_{1m} - i}{\frac{\sqrt{\sigma}}{m} - 1} + X \frac{\frac{\sqrt{\sigma}}{m}}{\frac{\sqrt{\sigma}}{m} - 1}. \tag{4.1}
 \end{aligned}$$

In this formula  $m = 0.92 \left( \frac{a}{b} \right)^2 + 0.002(90 - \beta_{2m})$ ;  $X$  is the adjusted experimental amount after consideration is given to the effect of a three-dimensional flow on the calculations.

#### B. Calculations using the airfoil equivalent flapping angle.

It is assumed in [10] that, if one takes the velocity triangle in the intake of the cascade on the stream surface and transforms it into an equivalent velocity triangle, then the conditions of the transformation are as follows: the radial coordinates for the meridional velocity in the exhaust are the same as those for the intake, and the value of  $C_{r'}$  is the same as the original exhaust velocity triangle. On the basis of these conditions, the equivalent exhaust aerodynamic angle is

$$\beta_{2me} = \operatorname{ctg}^{-1} \left[ \frac{u_1}{C_{m1}} - \frac{r_2}{r_1} \left( \frac{u_2}{C_{m1}} - \frac{C_{m2}}{C_{m1}} \operatorname{ctg} \beta_{2m} \right) \right], \tag{4.2}$$

The formula for the calculation of the lag angle is

$$\delta = \frac{\beta_{2me} - \beta_{1m} - i}{\frac{\sqrt{\sigma}}{m} - 1}. \tag{4.3}$$

(5) is the configuration of the projection on the plane of the cascade; moreover, in this planar surface, in order to find the equivalent exhaust flow angle

$$\beta_{2ze} = \operatorname{ctg}^{-1} \left[ \frac{u_1}{C_{z1}} - \frac{r_2}{r_1} \left( \frac{u_2}{C_{z1}} - \frac{C_{z2}}{C_{z1}} \operatorname{ctg} \beta_{2z} \right) \right], \tag{4.4}$$

when one is calculating the lag angle, one must make use of the

equivalent angle of flapping of the airfoil as it is projected on the plane of the cascade.

This program of calculations makes use of an airfoil lag angle which is calculated by using (4-1).

(3) The calculation of the neutral point on the suction face of the blade as well as the unique angle of incidence.

When one is designing an airfoil on the basis of the principle of a unique angle of incidence, then it is necessary to find the neutral point on the suction side of the blades involved. [2] points out that the use of the center point of all the start points for the Mach waves from the seals on the leading edges of the blades involved to the suction side of the blades gives a good approximation of the neutral point. This assumption has been verified by relatively detailed mapping out of the flows involved.

The calculations of this sequence are as follows. First, go through the calculation of the free flows and, in this way, obtain the distribution of streamline flows which correspond to Mach numbers  $M$  and flow angles  $\beta_m$ . After this, one can make use of the Prandtl-Meyer formula and figure out the Mach number when the angle of flow deviation from a free flow stream reaches a parallel state to the tangential direction at the point on the suction surface of the blade being considered. In this way, it is possible to obtain the distribution of the  $M$  numbers along the suction surfaces of the blades concerned and, from this, it is possible to find the neutral point. As far as three-dimentional flows are concerned, this sort of set up should cause the tangent line to the surface of the blade at the neutral point and the direction of free flow at that point to be the same; however, when one takes into consideration the blocking function of the blades involved as well as the boundary layers on these blades [2], then one can assume that there should be an angle

of incidence of  $\pm 1.5^\circ$  left or right between the tangent to the neutral point on a surface and the free flow at that point.

The angles of incidence for the leading edge mid-lines and the suction surfaces of blades as they were obtained by the use of the calculation methods outlined above are set out in Table 3. The data from [3] is also set out in the table.

#### (4) Shock wave position and shock wave losses

The model for the shock wave passage through the cascade is as shown in Figure 3. If we take a perpendicular to the mid-line of a passage at point A on the leading edge of a blade and drop it so that it intersects with the suction surface of another blade, then the perpendicular AB is a normal shock. In [1] and [3], the methods for figuring the M numbers in front of the shock waves were different. We chose to make use of the method from reference [3], that is, we found out the critical area ratio in front of the shock wave,  $(\frac{A}{A^*})_{sh}$ , and from this, we obtain the M number in front of the shock wave.

From the distribution of critical area ratios,  $(\frac{A}{A^*})_p$ , in free flows, we obtain a critical area ratio in front of the shock wave as follows:

$$\begin{aligned} \left(\frac{A}{A^*}\right)_{sh} &= \left(\frac{A}{A^*}\right)_p \frac{2r_m}{ts\sin\beta_m} \\ &= \left(\frac{A}{A^*}\right)_p \frac{r_m}{\pi r \sin\beta_m} \end{aligned} \quad (4.5)$$

In this formula,  $t$ ,  $r$ ,  $\beta_m$  are respectively the cascade distance which corresponds to point C, the radial coordinates and angle of free flow;  $r_m$  is the tangent radius inside the trough with C as its center.

The relationship between  $(\frac{A}{A^*})_{sh}$  and the Mach number in front of a shock wave,  $M_{sh}$ , is as follows:

$$\left(\frac{A}{A^*}\right)_{\infty} = \frac{1}{M_{\infty}} \left[ \left(\frac{2}{K+1}\right) \left(1 + \frac{K-1}{2} M_{\infty}^2\right) \right]^{\frac{K+1}{2(K-1)}}. \quad (4-6)$$

In Figure 3, the positions of the shock wave and the intersection point with the suction surface of the blade have definite requirements associated with them. In (3), it was required that the distance from a point near the vicinity of the tip of the blade to the forward edge of the blade be approximately  $1.25 b_f$ . With different radii, this numerical value will differ and one should, on the basis of this requirement, adjust the chord length ratio parameter,  $b/b_f$ , for multiple arc airfoils.

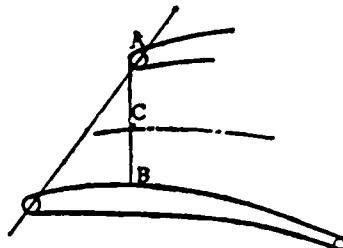


Figure 3

### (5) Checking on trough blockage

The cascade trough passage for the tangential plane to the conical surface is as shown in Figure 4. With the exception of a standard blade, in the opposite direction to the direction of the turning and, on the basis of the cascade distance of the conical surface, one gets the formation of the blade dish, and the blade cascade trough passage is formed from the back of the blade (the suction surface) of a standard blade. Due to changes in the cascade distance of the conical surfaces involved, the blade dishes will be different from those of straight standard blades.

Shock wave losses make use of the methods which have been discussed above for their computation. In the calculations of the leading and trailing edge flow fields of blades, one obtains the total loss coefficient  $\bar{w}_t$ . If one subtracts the shock wave loss coefficient,  $\bar{w}_s$ , from this total, then one gets the airfoil loss coefficient,  $\bar{w}_p$ . We may accept the fact that in front of the point at which the shock wave contacts the back of a blade (the suction surface) there are no losses. We can further assume that from that point to the trailing edge of the blade the values of the loss

coefficient of the airfoil are distributed in a linear way.

The computational formula is

$$\frac{A}{A^*} = \frac{m p_w^* \cdot 2 r_m \cos \varphi N_B}{\sqrt{T_w^*} \frac{\partial G_B}{\partial r}}, \quad (4.7)$$

$$p_w^* = p_{wL}^* \sigma \left( \frac{T_w^*}{T_{wL}^*} \right)^{\frac{K}{K-1}}, \quad (4.8)$$

$$T_w^* = T_{wL}^* \left\{ 1 + \frac{K-1}{2} \frac{r^2 w^2}{K g R T_{wL}^*} \left[ 1 - \left( \frac{r_L}{r} \right)^2 \right] \right\}, \quad (4.9)$$

$$\sigma = 1 - \frac{\bar{w}}{\left( \frac{T_w^*}{T_{wL}^*} \right)^{\frac{K}{K-1}}} \left[ 1 - \frac{1}{\left( 1 + \frac{K-1}{2} M_{wL}^2 \right)^{\frac{K}{K-1}}} \right], \quad (4.10)$$

$$\bar{w} = \bar{w}_s + \frac{\bar{w}_T - \bar{w}_s}{b - X_{sh}} (X - X_{sh}). \quad (4.11)$$

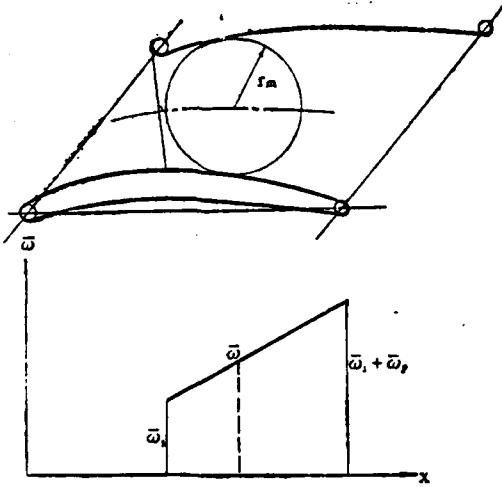


Figure 4

It is required that the smallest values of  $A/A^*$  be 1.02-1.05. If this range is exceeded, it is possible to adjust the camber ratio parameter for the airfoil  $F_s$ . The computations demonstrate that, in the vicinity of the blade tips, the influence which an adjustment of the camber ratio has on the smallest values of  $A/A^*$  is not great. In the vicinity of the base of a blade, the camber ratio does have an influence on the minimum values of  $A/A^*$ .

The results of the calculations are shown in Table 3.

## 5. CONCLUSIONS

Concerning the design methods and computer program sequences which have been presented in this article, it is possible, on the basis of given original conditions and design requirements, and by going through the three successive stages of computation--the calculation of flow fields for the leading and trailing edges of a blade, the approximate calculation of the  $S_1$  stream surface of revolution, and the configuration and successive integrations of the blades involved--it is possible to obtain the airfoil coordinates required for the manufacture of blades as well as the necessary data for computing strengths. Sample calculations clearly show that the results from these methods are quite good. Because of this, the methods which have been discussed in this article as well as the computational sequences presented have contributed an effective tool to the work of doing the aerodynamic designs for transsonic axial flow compressor stages as currently in use.

TABLE 1. A comparison of air flow parameters for  
the leading and trailing edges of blades

		0	10	40	80	100
Class I rotors	$\beta_{1m}$	1 46.54	41.72	34.48	28.18	25.34
		2 46.34	41.54	34.38	28.17	25.37
	$W_{1m}$	1 186.58	198.03	215.99	216.02	209.20
		2 185.28	195.64	215.94	216.09	209.58
	$\beta_{2m}$	1 110.09	84.99	52.13	36.84	34.20
		2 109.81	85.38	51.82	36.57	33.93
	$W_{2m}$	1 180.82	188.05	180.83	178.62	177.47
		2 183.87	191.54	182.70	177.82	175.88
Class I stators	$\alpha_{1m}$	1 35.30	4.2	47.15	50.45	48.83
		2 35.34	41.35	47.39	50.87	48.95
	$C_{1m}$	1 191.49	197.21	189.67	188.30	189.37
		2 192.06	198.14	190.68	190.14	191.84
	$\alpha_{2m}$	1 90	90	90	90	90
		2 90	90	90	90	90
	$C_{2m}$	1 190.18	197.94	201.43	201.31	201.68
		2 188.55	195.82	200.28	205.40	207.90

NOTE: Table 1 is a display of the results of computations based on the sequence put out in this article. Table 2 is a display of the data given in reference [3].

A comparison of air flow parameters  
 TABLE 2. for the leading and trailing edges of blades

Percentage of amount of flow		0	20	50	80	100
Rotors	$\beta_{1m}$	1	36.06	32.68	29.00	25.01
	$\beta_{1m}$	2	37.12	32.57	28.79	25.16
	$W_{1m}$	1	177.95	201.26	215.08	209.62
	$W_{1m}$	2	184.90	200.20	213.40	211.10
	$\beta_{2m}$	1	79.91	55.50	41.15	32.88
	$\beta_{2m}$	2	80.23	55.50	41.57	33.00
	$W_{2m}$	1	158.89	183.21	177.57	168.86
	$W_{2m}$	2	164.10	182.40	174.40	170.20
Stators	$\alpha_{1m}$	1	38.19	46.80	49.43	50.62
	$\alpha_{1m}$	2	38.49	46.92	49.59	50.55
	$C_{1m}$	1	193.40	212.00	210.54	213.55
	$C_{1m}$	2	195.50	213.00	212.00	213.80
	$\alpha_{2m}$	1	90	90	90	90
	$\alpha_{2m}$	2	90	90	90	90
	$C_{2m}$	1	183.16	214.48	210.40	205.44
	$C_{2m}$	2	187.00	214.60	211.30	208.90
201.18						
200.10						

NOTE: Table 1 is a display of the results of calculations done on the basis of the sequence put out in this article. Table 2 is a display of the data put out in reference (2).

TABLE 3. A comparison of calculated values for the parameters of the airfoils of a first stage rotor

<del>% of amt of para-meters</del> flow	0		41.45		80.70		100	
	1	2	1	2	1	2	1	2
$i_{\infty}$	-0.9	-0.9	1.2	1.2	2.58	2.80	3.72	3.60
$i_c$	3.84	3.90	4.08	3.48	3.91	4.20	4.83	4.70
$i_s$	/	/	1.72	1.70	1.50	1.50	1.50	1.50
$\delta$	15.83	15.90	7.53	7.80	7.00	6.80	11.74	11.00
$\bar{\omega}_s$	0	0	0.0415	0.0419	0.0741	0.0709	0.0899	0.0805
$(\frac{A}{A^*})_{\min}$	1.058	1.052	1.038	1.040	1.043	1.046	1.033	1.045
$\frac{\phi_f}{\phi}$	0.0565	0.0565	0.11	0.11	-0.0185	-0.0185	-0.135	-0.135
$\frac{b_f}{b}$	0.229	0.235	0.3975	0.3980	0.490	0.497	0.55	0.55
$\frac{a}{b}$	0.514	0.518	0.544	0.544	0.824	0.815	0.677	0.665
$\beta_{w1}$	45.64	45.40	35.88	35.80	30.73	31.00	29.06	29.00
$\beta_{1x}$	50.38	50.20	38.55	38.20	32.09	32.40	30.17	30.10
$\phi_f$	8.61	8.50	2.34	2.40	-0.212	-0.200	-2.16	-2.30
$b_f$	0.0210	0.0218	0.0421	0.0419	0.0544	0.0561	0.0636	0.0636

NOTE: Table 1 is a display of calculated values gotten from an application of the sequence put out in this article. Table 2 is a display of the values put out in reference (3).

A	mechanical equivalent of heat, trough passage area	$\phi$	airfoil deflection angle the included angle between the meridional plane streamlining and the Z axis
A/A*	ratio of trough area to critical area		
a	distance between point of greatest camber and leading edge	$\rho$	gas density
b	chord length	$\omega$	rotor angular velocity
c	absolute airflow speed	$\omega$	total pressure loss parameter
$F_s$	camber	$\sigma$	cascade density, entropy function
$G_B$	calculated amount of flow between calculation point and rotor hub	$\lambda$	included angle between calculation station 1 and the r axis
g	acceleration due to gravity	*	blockage parameter
H	total heat content	F	airfoil forward section
i	angle of incidence	r	free flow
K	index of absolute temperature	s	radial direction
L	curve calculation point	p	shock wave
m	meridional plane streamline, parameter used to calculate lag angle	sh	airfoil
M	Mach number	z	shock wave
N <sub>B</sub>	blade number	$\theta$	meridional plane streamline, cascade trough mid-line
p	pressure	c	axial direction
S	airflow entropy	ss	circumferential direction
R	gas constant	e	airfoil mid-line
r	radial coordinates	w	suction surface leading edge
T	temperature	L	equivalent speed triangle
t	cascade distance	B	corresponding parameter
u	circumferential speed	1	blade leading edge
W	corresponding speed	2	neutral point
X	corrected value for lag angle	0	blade leading edge airflow parameter
Z	axial coordinates		blade trailing edge airflow parameter
$\alpha$	stream surface conic angle		mechanical remote airflow parameter
$\beta$	corresponding airfoil angle		
$\gamma$	airfoil installation angle		
$\delta$	lag angle		

## REFERENCES

- [1] Wu, C. H., A General Theory of Three-Dimensional Flow in Subsonic and Supersonic Turbomachines of Axial-, Radial-, and Mixed-Flow Types, NACA TN 2604, 1952.
- [2] Monsarrat, N. T., Keenan, M. J., and Tramm, P. C., Design Report of Single-Stage Evaluation of Highly-Loaded High-Mach-Number Compressor Stage, NASA CR 72562, 1969.
- [3] Messenger, H. E., and Kennedy, E. E., Two-Stage Fan: I. Aerodynamic and Mechanical Design, NASA CR 120859, 1972.
- [4] Morris, A. L., Halle, J. E., and Kennedy, E. E., High-Loading 1800 ft/sec Tip Speed Transonic Compressor Fan Stage: I, Aerodynamic and Mechanical Design, NASA CR 120907, 1972.
- [5] Seyler, D. R., and Smith, L. H. Jr., Single Stage Experimental Evaluation of High Mach Number Compressor Rotor Blading: part I, Design of Rotor Blading, NASA CR 54581, 1967.
- [6] Gostelow, J. P., Krabacher, K. W., and Smith, L. H., Performance Comparisons of High Mach Number Compressor Rotor Blading, NASA CR 1256, 1968.
- [7] Sulam, D. H., Keenan, M. J., and Flynn, J. T., Single-Stage Evaluation of High-Loaded High-Mach-Number Compressor Stage: I, Data Performance, Multiple-Circular-Arc Rotor, NASA CR 72694, 1970.
- [8] Crouse, J. E., Janetzke, D. C., and Scbwirian, R. E., A Computer Program for Composing Compressor Blading from Simulated Circular-Arc Elements on Conical Surfaces, NASA TN D-5437, 1969.
- [9] Crouse, J. E., Computer Program for Definition of Transonic Axial-Flow Compressoer Blade Rows, NASA TN D-7345, 1974.

## Summary

# Analytical and Experimental Investigation of Performance of Supersonic Ejector Nozzle

*Shen Huili and Chen Zhongqing*

This paper presents an analytical method for calculating the flow field and performance of supersonic ejector nozzle. The calculations involve the real sonic line at the exit of the primary nozzle, the inviscid primary flow field, the correction for viscosity effect and the pumping, and thrust characteristics.

In order to bring calculated results into agreement with experimental data, the real sonic line, instead of the plane sonic line, is taken as the initial base line of calculation. The real sonic line is obtained by joining the points of intersection of constant flow angle lines in the throat region with Mach lines at the lip of the primary nozzle.

First, the inviscid primary flow field of the nozzle is calculated and then corrected to account for the viscosity effect. The method of correction for viscosity effect proposed in this paper replaces the original geometric coordinates of the ejector shroud with corrected geometric coordinates, which are obtained by superimposing on the original geometric coordinates the displacements of the mixed region and the boundary layer. On the basis of the corrected coordinates, the actual primary flow field and pumping performance of the nozzle are then calculated. The proposed method proves to be quite simple and accurate.

Calculations were performed on a "320" digital computer, and model tests on a ground test facility. The analytical and experimental results are found to be in fairly satisfactory agreement.

ANALYTICAL AND EXPERIMENTAL INVESTIGATION OF PERFORMANCE  
OF SUPERSONIC EJECTOR NOZZLE

Shen Huili and Chen Zhongqing

SUMMARY

This article introduces a theoretical calculation method to be applied to the performance and flow fields of supersonic ejector nozzles. It includes calculations such as these: the actual sonic line in the exhaust of a main jet tube, non-viscous primary flow fields, corrections for the influences of viscosity and pumping and thrust characteristics.

In order to make the results of calculations agree even better with data gathered through experimentation, we have made use of the real sonic line and not the planar sonic line as the initial base line for calculation purposes. The real sonic line is obtained by linking together the points of intersection between the constant flow angle lines in the throat region of the main jet nozzle and the Mach lines in the lip region of the main jet nozzle.

First of all, one must calculate the non-viscous primary flow field of the jet nozzle. After this is done, then one must consider necessary corrections for the effects of viscosity. The method which is used in this article for correcting to compensate for the effects of viscosity is to correct the geometric coordinates of the ejector shroud and replace the original coordinates with the corrected ones. These new coordinates are obtained by adding the displacement thicknesses of the mixed region and boundary layer to the original geometrical shroud coordinates. One can then compute the real jet nozzle flow field, pumping performance and other related characteristics on the basis of the corrected coordinates. This correction method is quite simple and adequately accurate.

Calculations were completed on a "320" digital computer and model tests were run at a ground facility. The theoretical and experimental results obtained were basically identical.

### FLOW GRAPHS [1, 2]

1 高次流流动(图1)

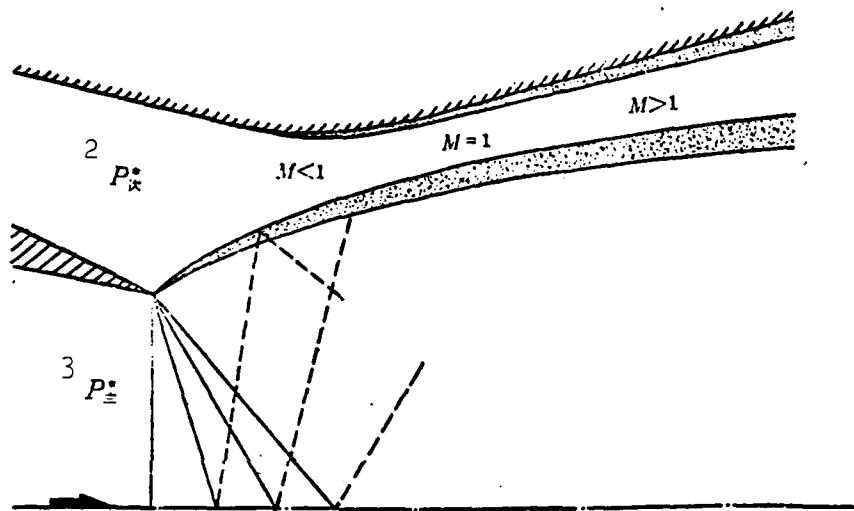


Figure 1. A flow chart for high harmonic flow in jet ejector nozzles.

Key: 1--higher harmonic flow (Figure 1); 2--harmonic;  
3--primary

When the amount of harmonic flow is relatively large, the primary flow continuously expands and increases in speed in the expansion section between the exhaust of the jet nozzle and the "fluid wall"; the harmonic flow then increases in speed in the passage formed by the shroud wall and the "fluid wall". On the basis of the dimensions of the ejector nozzle and aerodynamic conditions, there are three possible situations:

harmonic flow maintains a subsonic speed;

the harmonic flow is at sonic speed at the end of the jet nozzle exhaust;

the harmonic flow is supersonic at the end of the jet nozzle exhaust.

Due to the effect of gas viscosity, the area between the primary and harmonic flows forms a mixed layer of unequal pressures. The shroud wall forms a boundary layer. However, when one is dealing with a situation in which there is high harmonic flow, the influence of viscosity is relatively small.

Lower harmonic flow (Figure 2).

When the amount of harmonic flow is very small (or approaches zero), the primary flow takes the free flow form and expands it into an area of equal pressure. The viscosity of this jet adheres to the wall surface of the expansion section and, after a violent recompression, one gets the formation of a shock wave. In the upper reaches of this section, the low speed isobaric gases which are sealed in by the isobaric boundaries form a dead space. Along the boundaries of the main flow, due to the effects of the viscosity of the gases concerned, there is also a mixed region. This region separates the main flow and the dead space. In this mixed region, there is a momentum and energy transfer between the primary and harmonic flows. Due to the effects of viscosity, the primary flow continuously pulls out a certain amount of the harmonic flow from the dead space and the amount of the harmonic flow which is delayed in its flowing is then replaced by the system which supplies the harmonic flow.

#### METHOD OF CALCULATION

##### 1. Calculation of the real sonic line [3]

The real sonic line is determined after calculating the flow angles in the transsonic throat area of the main jet nozzle and the Mach lines in the lip area of the main jet nozzle (Figure 3). Let us assume that the airflow is stable, non-vortical, of equal entropy and two-dimensional. This type of flow can be described by using flow function equations, that is,

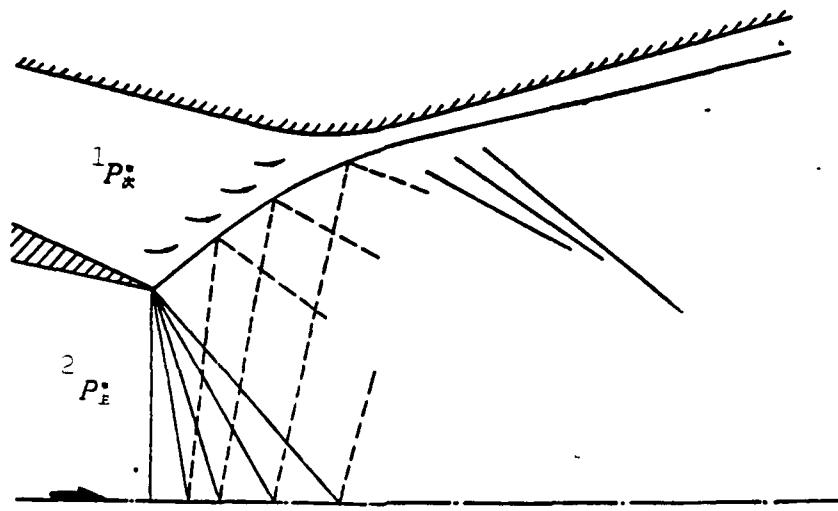


Figure 2. Chart of lower harmonic flow in jet ejector nozzle  
Key: 1--harmonic; 2--primary

$$\left[ 1 - \left( \frac{\rho_0}{\rho} \right)^2 \cdot \frac{\psi_y^2}{a^2} \right] \psi_{xx} + \left[ 2 \left( \frac{\rho_0}{\rho} \right)^2 \frac{\psi_x \psi_y}{a^2} \right] \psi_{xy} + \left[ 1 - \left( \frac{\rho_0}{\rho} \right)^2 \frac{\psi_x^2}{a^2} \right] \psi_{yy} = 0$$

In these equations, the main thing is to take the partial derivative of  $x$  or  $y$ . We can then substitute the speed variables defined below, that is,

$$q = (u^2 + v^2)^{\frac{1}{2}} \quad (2)$$

$$\theta = \arctg \left( \frac{v}{u} \right) \quad (3)$$

When we do this, then equation (1) becomes

$$\frac{\partial}{\partial q} \left( \frac{\rho_0}{\rho} q \frac{\partial \psi}{\partial q} \right) + \frac{\rho_0}{\rho} \frac{1}{q} \left( 1 - \frac{q^2}{a^2} \right) \frac{\partial^2 \psi}{\partial \theta^2} = 0. \quad (4)$$

If we resubstitute, then the definition becomes

$$d\omega = \frac{\rho}{\rho_0} \frac{dq}{q} \quad (5)$$

There is a transformation of speeds and the equation above simplifies to become

$$\frac{\partial^2 \psi}{\partial w^2} + K(M) \frac{\partial^2 \psi}{\partial \theta^2} = 0, \quad (6)$$

In this equation,

$$K(M) = \left(\frac{\rho_0}{\rho}\right)^2 (1 - M^2) \quad (7)$$

After substituting the "tangent gas assumption"  $K(M) = 1$ , it is possible to simplify equation (6). The reason for this is that the speed graph relationships can be simplified to the form of Cauchy-Riemann equations. Because of this fact, it is possible to solve for the flow fields involved by making use of complex variables. Compressible and non-compressible flows are linked by the equation

$$d\omega = \frac{\rho}{\rho_0} \frac{dq}{q}$$

When this is integrated, then we can obtain

$$\omega = \ln \frac{2 \left( \frac{q}{q_m} \right) \frac{M_m}{\sqrt{1 - M_m^2}}}{1 + \sqrt{1 + \frac{\left( \frac{q}{q_m} \right)^2 M_m^2}{1 - M_m^2}}}$$

In this equation,  $q_m$  and  $M_m$  represent the speed parameters and Mach numbers for matched states. Because of the fact that the conditions of the Cauchy-Riemann relationships are met and because the boundary conditions of the problems being studied are also satisfied, it is possible to make use of original and congruent complex functions in a set which satisfies the required boundary conditions, that is,

$$F(\omega - i\theta) = \ln \frac{\cosh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s) - \cosh \left( \frac{\pi}{\alpha} \Delta \omega \right)}{\cosh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s) - 1} \quad (9)$$

In this equation,  $\Delta \omega = \omega_s - \omega_s$ . If we carry out a differentiation of the equation above, then we obtain the complex velocity

$$\frac{dF}{d(\omega - i\theta)} = \frac{\frac{\pi}{\alpha} \sinh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s)}{\cosh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s) - \cosh \left( \frac{\pi}{\alpha} \Delta \omega \right)} \quad (10)$$

$$= \frac{\frac{\pi}{\alpha} \sinh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s)}{\cosh \frac{\pi}{\alpha}(\omega - i\theta - \omega_s) - 1} \quad (10)$$

After we make use of this of this solution to substitute into the complex variable  $z = x + iy$  as it pertains to the limited edges of a jet nozzle, the graphic solution for speed can be transferred to the physical plane. When this is done, then

$$dz = \frac{1}{e^{\omega-i\theta}} \cdot \frac{dF}{d(\omega-i\theta)} \cdot d(\omega-i\theta) - \frac{e^{-i\theta}}{4} \cdot \frac{dF}{d(\omega-i\theta)} \cdot d(\omega-i\theta) \quad (11)$$

In this equation one has the last quantity in the horizontal line representing the complex conjugate value. If one integrates equation (11), then it is possible to obtain the points  $(x, y)$  on the speed plane. The numerical constants in the equation can be adjusted in this way. Let the position of the lip of the exhaust of the jet nozzle on the speed plane,  $(\omega_j, \alpha)$ , correspond to the point  $(0,1)$  on the physical plane. In this sort of analysis, what is particularly useful for our purposes is the distribution of the angles of flow in the throat region of the nozzle being considered. Lines along which the angles of flow are equal are called "isogoral lines". In order to obtain the equation for an isogonal line, we must carry out a numerical value integration of the interval of flow angles from 0 to  $\alpha$  as against  $\omega$  from  $\omega$  to  $\omega_a$ .

We can use the characteristic line method in order to calculate the Mach line for the lip of the jet nozzle being considered. The main equations which are employed in this process are

$$\frac{dy}{dx} = \operatorname{tg}(\theta \pm \mu) \quad (12)$$

$$\frac{dp}{\rho q^2 \operatorname{tg} \mu} \pm d\theta + \frac{\sin \theta \cdot \sin \mu}{\sin(\theta \pm \mu)} \cdot \frac{dy}{y} = 0 \quad (13)$$

$$ds = 0 \quad \text{when } \frac{dy}{dx} = \operatorname{tg} \theta \text{ 时} \quad (14)$$

In these equations,  $\mu$  is the Mach angle for the place being considered. We choose the static pressure  $p$  and the flow angle  $\theta$  to be the basic variables. In the lip region of the jet nozzle being considered, the expansion takes the form of a central wave and it satisfies a Prandtl-Meyer function. When the flow through the

AD-A104 327

FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OH  
RECENT SELECTED PAPERS OF NORTHWESTERN POLYTECHNICAL UNIVERSITY—ETC(U)

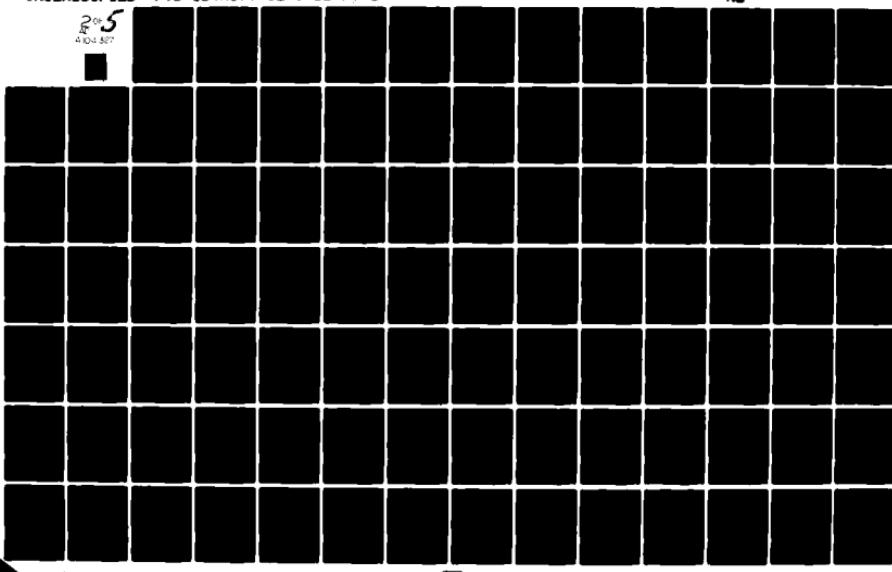
F/8 20/4

AUG 81

UNCLASSIFIED FTD-ID(RS)T-0259-81-PT-1

NL

205  
4104-327



lip region of a nozzle rotates through a certain angle, the expansion process can be divided into two finite numerical separation processes. The first point of the sonic line is the point of intersection between the Mach line of the lip region of the jet nozzle being studied and the isogonal line for the angle  $\alpha$  (Figure 3). The other points on the sonic line can be found using a similar method.

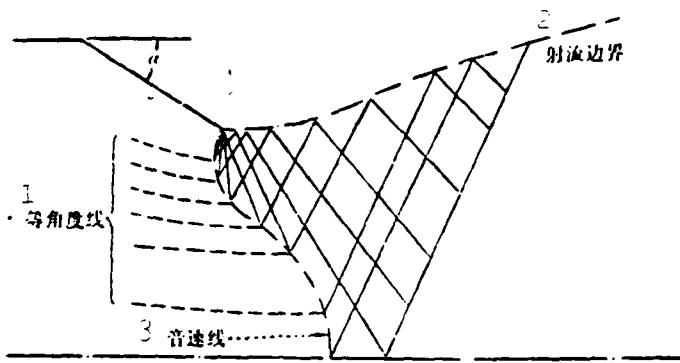


Figure 3. The actual sonic line

Key: 1--isogonal lines; 2--sonic line; 3--jet boundary

2. High harmonic non-viscous flow fields and characteristic calculations

(1) Stages

The harmonic flow is a one-dimensional equientropic flow; each set of corresponding points along the boundary between the primary and the harmonic flows is a point of equal static pressure; the primary flow is non-viscous and is controlled by the characteristic line.

(2) Conditions already known

Initial conditions of the primary flow--total pressure, total temperature and amount of flow; total pressure and total temperature of harmonic flow, external boundary atmospheric pressure; geometrical dimensions of ejector nozzle.

(3) The primary formulae utilized [6]:

$$G = mAq(\lambda) \frac{P^*}{\sqrt{T^*}} \quad (16)$$

$$\alpha = \text{Arc sin} \frac{1}{M} \quad (16)$$

$$\lambda^2 = 1 + \frac{2}{K-1} \sin^2 \left( \sqrt{\frac{K-1}{K+1}} \theta \right) \quad (17)$$

$$\frac{\delta p}{p} \frac{\sqrt{M^2-1}}{KM^2} + \delta\phi = - \frac{\sin\phi}{MR} \delta\eta \quad (18)$$

$$\frac{\delta p}{p} \frac{\sqrt{M^2-1}}{KM^2} - \delta\phi = - \frac{\sin\phi}{MR} \delta\xi \quad (19)$$

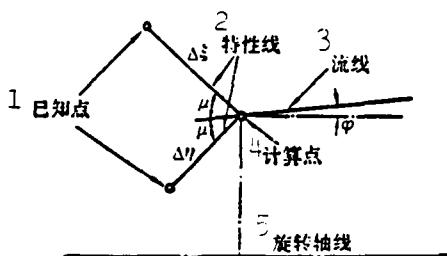


Figure 4. Calculation graph for ejector nozzle

Key: 1--already known point; 2--characteristic line; 3--streamline; 4--calculation point; 5--line of rotational axis

(4) Calculation procedure. First, decide on a certain amount of harmonic flow. On the basis of the condition that the static pressures of the primary and harmonic flows are equal, determine the expansion which the exhaust of the main jet nozzle handles. The determination of the points on the boundary between the two airflows is done by the use of the successive substitution method. As far as the static pressure of the harmonic flow is concerned, one writes the continuity equations for the harmonic flow between two different cross-sections of it. In connection with another aspect of the problem, one writes the equation for the characteristic lines which appear in the vicinity of these points. These calculations are carried right out to the jet nozzle

exhaust. If one is dealing with a case in which the speed of the harmonic flow is subsonic, then the determined condition is that the static pressure of the harmonic flow in the exhaust of the jet nozzle is equal to the surrounding atmospheric pressure. When the harmonic flow is supersonic in the plane of the exhaust, then the harmonic flow has a throat section in which the M number for that cross-section is 1. At such a time, the determined condition is that the amount of flow which was chosen and the amount of flow which can go through the throat area of the nozzle are the same.

### 3. Correction for the influence of viscosity [4,3,7]

In its consideration of the influence of viscosity, reference [4] is concerned with the form of the cross-sectional area of the harmonic flow through the throat region and the correction of it. Reference [3] then deals with the correction of the boundary lines of the primary and harmonic flows and considers the influences of the boundary layers on the outside walls. The first method is simple; however, the accuracy of it is somewhat less than excellent. The latter method for correction is relatively complicated. The method which this article puts forward is the external shroud coordinate correction method. The procedure for this method is as follows:

- (1) first, on the basis of the dimensions of the original external shroud, calculate the non-viscous flow field;
- (2) calculate the displacement thicknesses of the mixture layers of the primary and harmonic flows;
- (3) calculate the displacement thickness of the boundary layer on the interior walls of the shroud;
- (4) subtract the displacement thicknesses of the boundary layers from the dimensions of the passage cross-sections for the harmonic flow as these were obtained from the calculations of the non-viscous flow field. Then, add the displacement thickness of the mixture layer, that is to say, the dimensions of the passage

cross-section of the harmonic flow as it was obtained after correction;

(5) on the basis of the shroud wall coordinates after correction, reutilize the calculation method which was used to figure the non-viscous flow field in order to figure out the entire flow field. The details of these calculations are as presented below.

(1) the calculation of the displacement thickness  $\delta_1^*$ , of the mixed layer harmonic flow.

If we recognize the fact that the mixed layer is an isobaric one, then the speed graph for the interior of the mixed layer (Figure 5) will be capable of being obtained from the various equations presented below [4]:

$$\varphi = \frac{u}{u_a} = \frac{1 + \varphi_b}{2} + \frac{1 - \varphi_b}{2} \left( \frac{1}{\pi^{\frac{1}{2}}} \int_0^y e^{-\beta^2} \cdot d\beta \right) = \frac{1 + \varphi_b}{2} + \frac{1 - \varphi_b}{2} \operatorname{erf} \eta \quad (20)$$

In these equations,  $\varphi_b = \frac{u_b}{u_a}$ ,  $u_b$ —harmonic flow speed,  $u_a$ —main flow speed;  $u$ —velocity component in the  $x$  direction

$\operatorname{erf} \eta = \frac{2}{\pi^{\frac{1}{2}}} \int_0^y e^{-\beta^2} \cdot d\beta$ ,  $\eta = \sigma \frac{y}{x}$ ,  $\sigma$ —similarity parameter obtained from a semi-empirical formula;

$y, x$ —the coordinates in the coordinate system of the interior of the mixed region, the air-flow speed at the point where  $y=0$

is  $u = \frac{u_a + u_b}{2}$  (that is  $\varphi = \frac{1 + \varphi_b}{2}$ ),

$\eta$  is a similar coordinate, that is to say, that it is only necessary for the value of  $y/x$  to be the same, and the value of  $u/u_a$  will also be the same. To say it another way, the speed graphs for the various cross-sections of the mixed region are all similar.

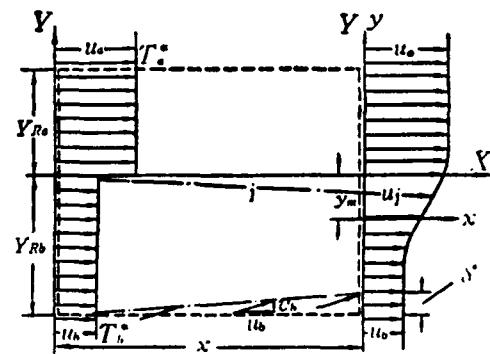


Figure 5. The speed chart for the interior of the mixed layer.

If we take the Prandtl number for disturbed flow to be 1, ignore  $\frac{\partial p}{\partial z}$ , and assume that  $c_p$  is a constant, then it is possible to obtain the temperature graph directly from the speed graphs, that is to say,

$$A = \frac{T^*}{T_a^*} = \frac{1}{1-\varphi_b} \left[ \left( \frac{T_a^*}{T_b^*} - \varphi_b \right) + \left( 1 - \frac{T_b^*}{T_a^*} \right) \varphi_b \right] \quad (21)$$

In this equation

$T^*$ --the total temperature inside the mixed region

$T_a^*$ --the total temperature of the main flow

$T_b^*$ --the total temperature of the harmonic flow.

We already know the speeds for the various points in the interior of the mixed layer, the total temperature and the static pressure (equal to the static pressures of the main and harmonic flows on the boundaries of the mixed layer). Knowing this, we can then solve for the displacement thickness of the mixed layer.

The speed graphs and temperature graphs of the mixed layer, which we talked about above, are all given in terms of the coordinates  $x, y$  of the internal coordinate system. The relationship between these coordinates and the coordinates,  $X, Y$ , which are found in the coordinate systems used in the reference for this article, is

$$Y = y - y_m$$

The jet boundary lines of the mixed layer use  $j$  for their representation (see Figure 5). The point at which  $y=y_j$  is appropriate to take as the point at which the mixed layer is divided into two parts. From  $\eta$  to  $\eta_{re}$ , the amount of flow which goes through this range can be exactly equated with the amount of flow which enters the mixed layer from the main flow. From  $\eta$  to  $\eta_{re}$  the amount of flow which goes through this range is just equal to the amount of flow which enters the mixed layer from the harmonic flow.

On the basis of the  $\eta$ , which is solved by the use of the principle of the conservation of momentum and mass both before and after mixing,

$$I_1(\eta_i) = [I_1(\eta_{Ra}) - I_1(\eta_{Ra})]/(1 - \varphi_b) \quad (22)$$

$$I_1(\eta) = \frac{(1 - c_s^2)\varphi_b \eta_{Ra}}{(T_b^* / T_a^*) - c_s^2 \varphi_b^2} + \int_{\eta_{Ra}}^{\eta} \frac{(1 - c_s^2)\varphi}{1 - c_s^2 \varphi^2} d\eta \quad (23)$$

$$I_2(\eta) = \frac{(1 - c_s^2)\varphi_b^2 \eta_{Ra}}{(T_b^* / T_a^*) - c_s^2 \varphi_b^2} + \int_{\eta_{Ra}}^{\eta} \frac{(1 - c_s^2)\varphi^2}{1 - c_s^2 \varphi^2} d\eta \quad (24)$$

In these equations,  $c$  is the Clark number and

$$c^2 = \frac{M^2}{\left(\frac{2}{K-1}\right) + M^2} \quad (25)$$

$\eta_m$  is also solved for by the use of the principle of the conservation of momentum and mass. If one uses similar coordinates to give this, then

$$\eta_m = \frac{\sigma y_m}{x} = \eta_{Ra} - \frac{1}{1 - \varphi_b} [I_2(\eta_{Ra}) - \varphi_b I_1(\eta_{Ra})]$$

On the basis of the concept of the displacement thickness of the boundary layers, the displacement thickness  $\delta^*$  mixing which is given for the mixing layer is

$$\delta^* \text{ mixing} = \frac{\sigma v_b}{x} = \frac{\left(\frac{T_b^*}{T_a^*}\right) - c_s^2 \varphi_b^2}{(1 - c_s^2) \varphi_b} I_1(\eta_i) - \eta_m, \quad (26)$$

In this,

$$\sigma = \left[ \frac{\sigma}{\sigma_1} \right] \cdot \sigma_1, \text{ and}$$

$$\sigma_1 = 12 + \frac{2.758 \cdot C_{s1}}{\sqrt{\frac{K_{s1}-1}{(1-C_{s1}^2)^{\frac{K_{s1}-1}{2}}}}}, \quad (27)$$

$$K_{s1} = \frac{\frac{G_{\pm} \sqrt{T_b^*}}{G_{\pm} \sqrt{T_a^*}}}{1 + \frac{G_{\pm} \sqrt{T_b^*}}{G_{\pm} \sqrt{T_a^*}}}, \quad (28)$$

(1--harmonic 2--main)

$$\frac{\sigma}{\sigma_1} = \frac{1 + \varphi_b}{1 - \varphi_b} \quad (30)$$

$$C_{s1} = \frac{C_s^2 (1 - \varphi_b^2)}{C_s^2 (1 - \varphi_b)^2 + (1 - C_s^2)}$$

(2) The calculation of the displacement thickness  $\delta^*$  boundary of the boundary layers on the interior walls of the exterior shroud.

Concerning the matter of using the displacement thickness,  $\delta^*$  boundary to calculate the influence which the boundary layers on the interior walls of the exterior shroud have on the amount of the harmonic flow, the properties of the boundary layers are calculated on a base of equivalent length and the equivalent length is defined to be

$$X = (Gy^*)^{-1} \int_0^x (Gy^*) dx \quad (31)$$

In this equation

$$G = \left( \frac{M}{1 + 0.2M^2} \right)^{\frac{1}{4}}$$

$\beta$  can be taken to be 1.20 or 1.25; the choice of its value depends on the Reynolds number. The Reynolds number of the place where the equivalent length for the solution of the equation above is located,  $R_s$ , is calculated with the use of total pressure and total temperature, that is to say,

$$R_s = \frac{\sqrt{\frac{kgRT^*}{\gamma^*}}}{M} (1 + 0.2M^2)^{-0.25} \quad (32)$$

In this equation,

$$\gamma^* = 9.81T^*R \frac{\mu(T^*)}{p^*} \quad (33)$$

Concerning the matter of using the Reynolds number for a certain place,  $R_s$  in order to calculate the displacement thickness,  $\delta^*$  boundary, and the momentum thickness  $\theta$ , when the order of  $R_x$  is  $10^5$ , one can use the equations below:

$$\delta^* \text{ boundary} = 0.046 \cdot x \cdot (1 + 0.8M^2)^{0.44} \cdot R_s^{-0.20} \quad (34)$$

$$\theta = 0.036 \cdot x \cdot (1 + 0.1M^2)^{-0.70} \cdot R_s^{-0.20} \quad (35)$$

When the order of  $R_x$  is  $10^7$ , then one can use the equations below:

$$\delta^* \text{ boundary} = 0.028 \cdot x \cdot (1.0 + 0.8M^2)^{0.44} \cdot R_s^{-0.167} \quad (36)$$

$$\theta = 0.022 \cdot x \cdot (1.0 + 0.1M^2)^{-0.70} \cdot R_s^{-0.167} \quad (37)$$

The equations above are accurate for  $K=1.4$ .

Due to the fact that we recognize the static pressure in the boundary layers to be constant, the pressure which we calculate off the surface,  $\delta^*$ , can be taken as being the surface pressure. This pressure performs the function of an iterative base. Because of this, the calculations can be carried out repeatedly on a standard form which is the same as being on the surface.

#### 4. Thrust performance calculations [5]

The thrust of the engine is given by the relationship set out below:

$$R_g = \frac{G}{g} V_e + p_e A_e - p_{in} A_e - \frac{G}{g} V_0 = R_a - \frac{G}{g} V_0 \quad (35)$$

$R_g$  in equation (38) is the total thrust, that is

$$R_a = F_a + F_b + R_{sh} - p_{in} A_e \quad (36)$$

On the basis of the principles of momentum, the equation above can be changed to become

$$R_a = F_a + F_b + R_{sh} - p_{in} A_e \quad (37)$$

and

$$F_a = \pi(\lambda_{ac})(1.0 + K_a C_D \cdot C_v \cdot M_{ac}^2) A_a \cdot p_{in}^2 \quad (\text{when } M_{ac} = 1) \quad (38)$$

$$F_b = p_e \cdot A_{ac} (1.0 + K_b \cdot M_{ac}^2) C_D \quad (39)$$

$$R_{sh} = \int_0^L (p_b \cdot 2\pi y \cdot \operatorname{tg}(\phi_{sh}) + C_f \cdot \pi \cdot y \cdot K_b \cdot p_b \cdot M_b^2 \sqrt{1 + \operatorname{tg}^2(\phi_{sh})}) dx \quad (40)$$

In these equations

$C_f$  -- coefficient of friction of the wall surface of the outside shroud

$C_d$  -- coefficient of the amount of flow in the main jet nozzle

$C_v$  -- the coefficient of velocity in the main jet nozzle

$y$  -- the radius of the wall surface of the exterior shroud

When one is not considering the influence of viscosity, then the coefficient of the amount of flow of the main jet nozzle and the speed coefficient can be respectively calculated using the equations below, that is

$$C_D = \frac{\int 2\pi y (\cos \phi dy - \sin \phi dx)}{\int 2\pi y dy} \quad (44)$$

$$C_V = \frac{\int 2\pi y (\cos \phi dy - \sin \phi dx) \cos \phi}{\int 2\pi y (\cos \phi dy - \sin \phi dx)} \quad (45)$$

In these equations, integration is carried out over the sonic line.

When one is considering the influence of viscosity, then the coefficient for the amount of flow of the main jet nozzle as well as the coefficient of velocity can be figured by using the formulae set out below:

$$C_D = (1.0 - K_1 \cdot R_e^{-0.20}) \frac{\int 2\pi y (\cos \phi dy - \sin \phi dx)}{\int 2\pi y dy}$$

$$C_V = (1.0 - K_2 \cdot R_e^{-0.20}) \frac{\int 2\pi y (\cos \phi dy - \sin \phi dx) \cos \phi}{\int 2\pi y (\cos \phi dy - \sin \phi dx)}$$

In these equations,

$$K_1 = \frac{0.185}{\cos |\phi_c|}, \quad K_2 = 0.144.$$

Also, in these equations

$\phi_c$  -- the semipyramid angle of the main jet nozzle

$\phi$  -- the included angle between the airflow speed on the sonic line and the x axis

$y$  -- the vertical coordinate of points on the sonic line

$x$  -- the horizontal coordinate of points on the sonic line

$R_e$  -- the Reynolds number

#### A COMPARISON OF THEORETICAL CALCULATIONS AND EXPERIMENTAL RESULTS

Concerning the use of the calculation method and sequence put forward in this article, we have used them to calculate the flow field of the jet ejector nozzle, the pressure distribution of the wall surfaces and the pumping performance of the jet nozzle. Moreover, we also carried out a comparison with the experimental

results obtained by testing with a model of the jet nozzle. The results were as shown below:

1. The flow field of the jet ejector nozzle and the pressure distribution of the wall surfaces.

The input data were:

$NW=18$  the number of functional nodal points in the coordinate system of the shroud

$N2=20$  the number of points picked on the sonic line

$N1=29$  the number of expansion waves picked to make the characteristic curve net

$N4=100$  the termination number of the cyclical variable J

$KW=0$  the x coordinate of the shroud in the exhaust cross-section of the main jet nozzle

$KC=0$  the x coordinate on the streamlines of the exhaust cross-section of the main jet nozzle

$YC=0.045m$  the radius of the main jet nozzle

$V=0.98 \text{ kg/sec}$  the initial value of the amount of flow of the harmonic flow

$G_{\text{main}}=14.4258 \text{ kg/sec}$  amount of flow in the main flow

$p_b^*=30702 \text{ kg/m}^2$  the total pressure of the main flow

$T_b^*=288^\circ\text{K}$  the total temperature of the harmonic flow

$K_b=1.4$  the specific heat of the harmonic flow

$R_b=29.27 \text{ kg-m/kg}^\circ\text{K}$  the gas constant of the main flow

$T_a^*=288^\circ\text{K}$  the total temperature of the main flow

$P_a^*=97992 \text{ kg/sec}^2$  the total pressure of the main flow

$R_a=29.27 \text{ kg-m/kg}^\circ\text{K}$  the gas constant of the main flow

$X_r=0.090m$  the distance from the exhaust of the main jet nozzle to the exhaust of the shroud

$K_a=1.4$  the specific heat of the main flow

$p_H=10333 \text{ kg/m}^2$  the atmospheric pressure on the outside boundary

The shroud turbine gallery curve uses a numerical set (list function) type of input.

$XS = X/YC$	$WS = Y/YC$	$XS = X/YC$	$WS = Y/YC$
0	1.2780	0.6889	1.1289
0.1111	1.2458	0.7556	1.1333
0.2222	1.2133	0.8222	1.1424
0.3333	1.1809	0.8889	1.1542
0.3778	1.1680	1.1111	1.1933
0.4444	1.1511	1.3333	1.2324
0.4889	1.1411	1.5556	1.2710
0.5556	1.1344	1.7780	1.3107
0.6222	1.1267	2.0000	1.3500

The output is the pressures,  $p$ , for all the various nodal points on the characteristic curve network, the coordinates in  $x, y$  as well as the angle of flow  $\phi$ ; it also includes the amount of flow of the harmonic flow,  $GS$ , as well as other required parameters.

The characteristic curve which was obtained by calculation for the jet ejector nozzle is as shown in Figure 6.

The calculated values for the pressure distribution of the wall surfaces of the shroud as well as the experimental results for the same variable are as shown in Figure 7.

## 2. Pumping characteristics of the jet ejector nozzle

What Figure 8 shows is this. Curve 1 is a display of the results of calculations of the non-viscous flow. Curve 2 is the result after the corrections had been made for viscosity. Figure 8 points out that after viscosity has been corrected for, the theoretical calculations and the experimental results are basically the same.

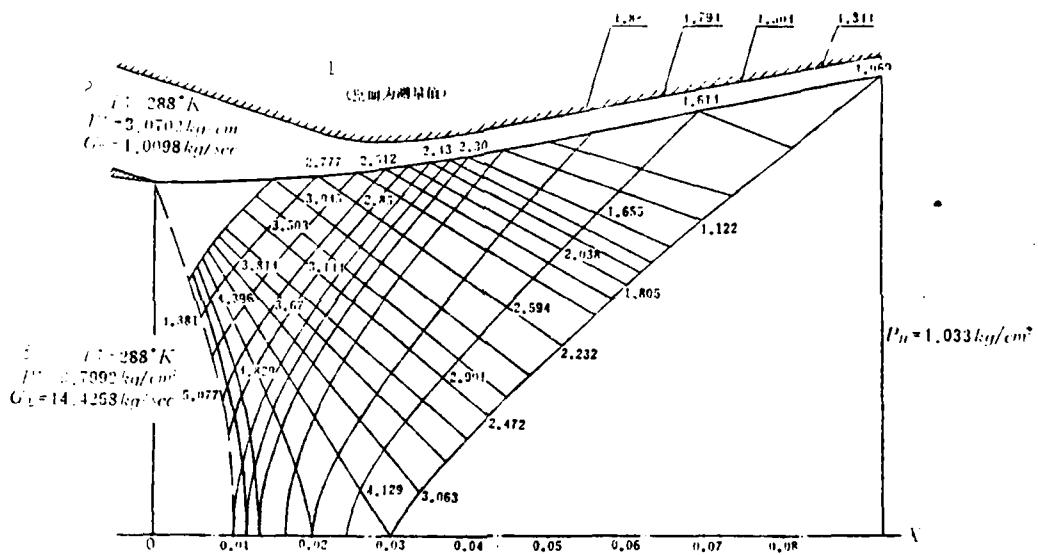


Figure 6. A graph of the characteristic curves of the jet-ejector nozzle.  
Key: 1--the wall surface is the measurement value; 2--harmonic; 3--main

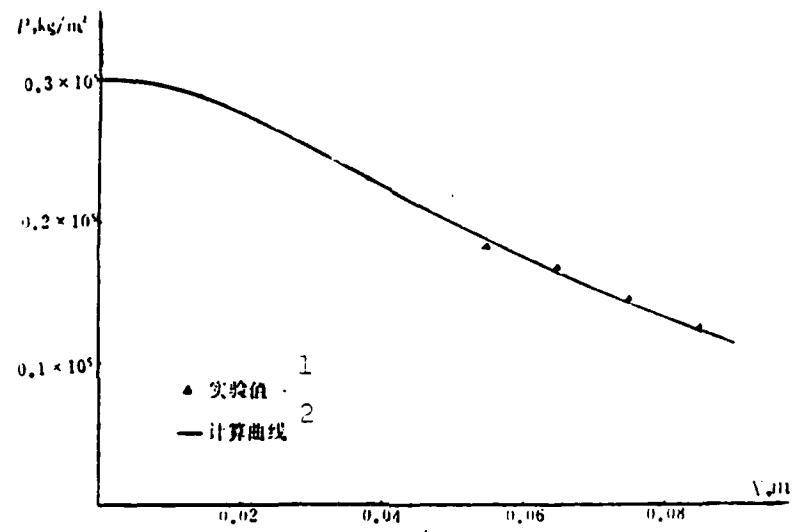


Figure 7. The wall surface pressure distribution for the shroud  
 Key: 1--experimental value; 2--computed curve

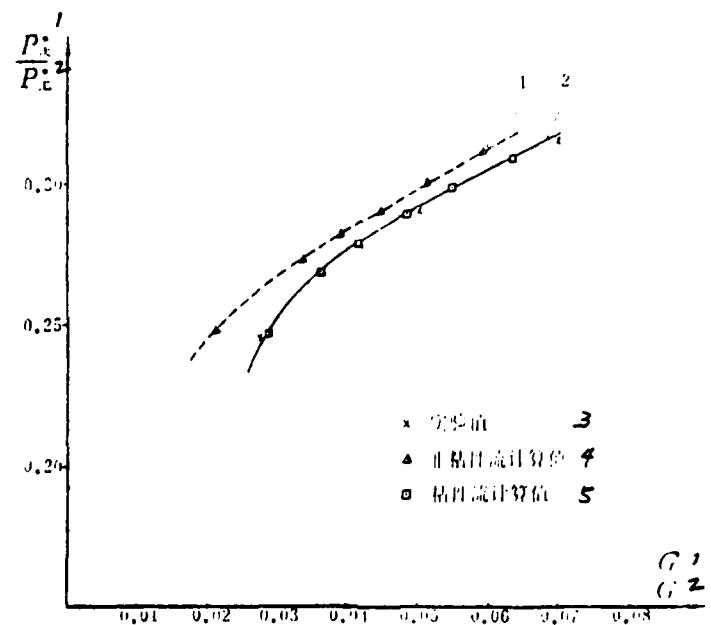


Figure 8. Pumping performance of the jet ejector nozzle.  
 Key: 1--main; 2--harmonic; 3--experimental value; 4--value computed on the assumption of non-viscosity; 5--value computed on the assumption of viscous flow

Figure 8 also shows that when one is operating with a high harmonic flow situation, the degree of error between the results of calculations of the non-viscous flow and the experimental data does not exceed 10%.

#### CONCLUSIONS

1. The methods of calculation which have been introduced and the results from experimentation agree closely enough.
2. The method which this article introduces for the correction to reflect the effects of viscosity is relatively simple and has the necessary accuracy.

TABLE OF SYMBOLS

$\psi$	flow function	K	specific heat
$\rho$	density	R	thrust or gas constant or radius
$a$	speed of sound	A	area
$u$	axial speed	F	impulse or function relationship
$v$	radial speed or the flow at the exhaust of the engine	$\eta$	Mach line coordinate
$p$	pressure	$\xi$	Mach line coordinate
$\mu$	Mach angle	$\beta$	chosen on the basis of the Reynolds number constant
$\phi$	flow angle (the angle between the streamline and the horizontal axis)	$\gamma^*$	viscosity coefficient used to express the blockage coefficient
$\theta$	the angle included between the characteristic line and the vertical axis or the boundary layer momentum thickness	$\delta^*$	boundary layer displacement thickness
$X$	equivalent length	subscripts <sub>2</sub>	
$y$	radial distance divided by the radius of the jet nozzle	$a$	main a--of the main flow
$x$	axial distance divided by the radius	$b$	harmonic b--of the harmonic flow
$\omega$	transformation speed	C	exhaust of the main jet nozzle
$\alpha$	semipyramid angle of the exhaust of the main jet nozzle	e	cross-section of the exhaust of the engine
$T$	temperature	H	boundary conditions
$G$	amount of flow	G	total
$M$	Mach number	O	intake of the engine
$\lambda$	coefficient of flow speed	SH	exterior shroud
		superscripts <sub>1</sub>	*--blockage parameter

## REFERENCES

- [1] Hardy, J. M., *Etude Théorique d'une Turière Convergente-Divergente Bi-Flux*, L'Aéronautique et l'Astronautique, No. 37, 1972-5, pp. 23-37.
- [2] Hardy, J. M., et Lacombe, H., *Les Tuyères Supersoniques à Double Flux. Méthodes de Calcul*, Revue Française de Mécanique, No. 24, 4 trimestre, 1967, pp. 49-59.
- [3] Anderson, B. H., *Computer Program for Calculating the Field of Supersonic Ejector nozzles*, NASA TND-7601, 1974, pp 1-86.
- [4] Chow, W. L., and Addy, A. L., *Interaction between Primary and Secondary Streams of Supersonic Ejector Systems and Their Performance Characteristics*, AIAA Journal, Vol. 2, No. 4, April 1964, pp 686-695.
- [5] Korst, H. H., Addy, A. V., and Chow, W. L., *Installed Performance of Air-Augmented Nozzles Based on Analytical Determination of Internal Ejector Characteristics*, Journal of Aircraft, Vol. 3, No. 6, Nov.-Dec. 1966, pp. 498-506.
- [6] Carriere, P., *Exhaust Nozzles. Supersonic Turbo-jet propulsion Systems and Components*, AGARD graph No. 120, 1969, pp. 287-374.
- [7] Anderson, B. H., *Factors which Influence the Analysis and Design of Ejector Nozzles*, AIAA, Paper 72-46, Jan. 1972.

## Summary

# An Optimum Design Procedure of Total-Temperature Thermocouple Probes

*Liu Cihong and Zhao Jueliang*

This paper presents a procedure for the design of total-temperature thermocouple probes to achieve minimum steady-state error under given operating conditions. First, the gas velocity in the sheath, i. e., the optimum internal flow velocity is determined on the principle that the sum of the radiation error, the conduction error and the velocity error is a minimum under given operating conditions. After determining the optimum internal flow velocity, the diameter of the exhaust hole is then determined and the layout of the probe is designed. This paper gives a brief presentation including: the programming of the optimum design of total-temperature thermocouple probes, the formula for manually estimating the approximate optimum internal flow velocity, and the design procedure. In the computation of heat conduction error, the heat-conduction equation of two thin rods or the average conductivity is used, then the thermal conductivities of the two wires are further taken into account. Besides, in calculating the optimum internal flow velocity, the heat transfer as the gas flows from the inlet of the sheath to the thermocouple junction is taken into account by restricting the internal flow velocity to be no less than the minimum internal flow velocity.

The design of the turbojet exhaust gas thermocouple probe and its verification tests reveal that, under given operating conditions, the error of the thermocouple probe reaches a minimum when the diameter of the exhaust hole is 1.4 mm. The thermocouple probe error  $\sigma_T$  and recovery factor  $r$  measured in the hot wind tunnel and the calibration wind tunnel differ negligibly from theoretically computed results as follows:  $\Delta\sigma_T=0.5^\circ\text{C}$ ,  $\Delta\sigma_T/\sigma_T=0.06\%$ ,  $\Delta r=0.005$ ,  $\Delta r/r=0.5\%$ . This appears to prove that the design procedure presented in this paper is quite satisfactory.

The optimum design procedure of the total-temperature thermocouple probe not only points out the highest accuracy that may be achieved under given operating conditions, but also shows how to make the minimum-error thermocouple probe a reality. It does not require the designers to have a great deal of experience and also allows greater machining tolerance for the exhaust hole.

AN OPTIMUM DESIGN PROCEDURE OF TOTAL TEMPERATURE  
THERMOCOUPLE PROBES

by

Liu Cihong and Zhao Jue lang

# An Optimum Design Procedure of Total-Temperature Thermocouple Probes

*Liu Cihong and Zhao Jueliang*

This paper presents a procedure for the design of total-temperature thermocouple probes to achieve minimum steady-state error under given operating conditions. First, the gas velocity in the sheath, i. e., the optimum internal flow velocity is determined on the principle that the sum of the radiation error, the conduction error and the velocity error is a minimum under given operating conditions. After determining the optimum internal flow velocity, the diameter of the exhaust hole is then determined and the layout of the probe is designed. This paper gives a brief presentation including: the programming of the optimum design of total-temperature thermocouple probes, the formula for manually estimating the approximate optimum internal flow velocity, and the design procedure. In the computation of heat conduction error, the heat-conduction equation of two thin rods or the average conductivity is used, then the thermal conductivities of the two wires are further taken into account. Besides, in calculating the optimum internal flow velocity, the heat transfer as the gas flows from the inlet of the sheath to the thermocouple junction is taken into account by restricting the internal flow velocity to be no less than the minimum internal flow velocity.

The design of the turbojet exhaust gas thermocouple probe and its verification tests reveal that, under given operating conditions, the error of the thermocouple probe reaches a minimum when the diameter of the exhaust hole is 1.4 mm. The thermocouple probe error  $\sigma_r$  and recovery factor  $r$  measured in the hot wind tunnel and the calibration wind tunnel differ negligibly from theoretically computed results as follows:  $\Delta\sigma_r = 0.5^\circ\text{C}$ ,  $\Delta\sigma_r/\sigma_r = 0.06\%$ ,  $\Delta r = 0.005$ ,  $\Delta r/r = 0.5\%$ . This appears to prove that the design procedure presented in this paper is quite satisfactory.

The optimum design procedure of the total-temperature thermocouple probe not only points out the highest accuracy that may be achieved under given operating conditions, but also shows how to make the minimum-error thermocouple probe a reality. It does not require the designers to have a great deal of experience and also allows greater machining tolerance for the exhaust hole.

AN OPTIMUM DESIGN PROCEDURE OF TOTAL TEMPERATURE  
THERMOCOUPLE PROBES

by

Liu Chihong and Zhao Jue liang

I. INTRODUCTION

Along with the advancement in jet engine and missile technology, the measurement of gas flow temperature becomes more and more important. One of the means to determine the temperature of gas flows is to use total temperature thermocouple probes to directly obtain the total temperature with the accuracy required. However, empirically designed total temperature thermocouple probes usually do not guarantee the precision demanded. Haig proposed a design procedure which enables us to obtain the accuracy required, but it still relies on an empirical method to distribute the thermocouples. It is very difficult to be absolutely sure that the thermocouple is at its best. This paper introduces an optimum design procedure of total temperature thermocouple probes based on the determination of the optimum internal flow velocity in order to overcome the shortcomings of Haig's method. The thermocouple remains in the best condition with minimum equilibrium error.

Under equilibrium measurement conditions, due to the loss of the kinetic energy of the gas flow and the heat transfer between the thermocouple and its environment, the total temperature value indicated by the thermocouple probe is deviating from the total temperature of the gas flow showing velocity error, radiation error and conduction error. The design of total temperature probes is to control these errors within a tolerable range using some kind of a shield for a specific application. Other sources of error such as wire materials,

circuits and instruments are not directly relevant to the design of total temperature thermocouple probes, therefore, they will not be considered here.

### 1. Velocity Error

Velocity error of a thermocouple is:

$$\sigma_v = (1-r) A \cdot \frac{W^2}{2gC_p} = (1-r) \left[ \frac{\frac{K-1}{2} M^2}{1 + \frac{K-1}{2} M^2} \right] T^* \quad (1)$$

where  $r$  is the complex thermal coefficient at the measurement point,  $M$  is the Mach number at the measurement point; if  $r$  represents the complex thermal coefficient of the whole thermal couple in the sheath then  $M$  is the Mach number of the incoming flow.

To minimize velocity error, the major means is to install a flow retarding device to decrease the velocity at the measuring point which enables the thermocouple wires to remain parallel with the gas flow.

In order to avoid the effect of contraction of the streamline near the inlet of the sheath and to obtain better shielding effects, measurements are usually taken at a point at a distance at least  $2 \sim 4$  times the internal radius of the sheath away from the inlet. During the flow of gas from the inlet to the thermocouple junction, heat transfer occurs because the insulated internal wall of the sheath is not isothermal. Consequently, the temperature of the gas flow at the junction point is lowered. The value indicated by the thermocouple will be low.

In practice, it is attributed to the decrease in complex thermal coefficient. The lower the internal velocity, the longer the holding time of the gas flow in the sheath, the larger the decrease in corresponding complex thermal coefficient becomes. Therefore, with respect to velocity error it is not

necessarily more advantageous to have a very small internal velocity. There is a lower limit in this case. Usually layer flow exists in the sheath and the minimum internal flow velocity can be calculated using the equation below:

$$u_{\min} = 0.545 w^{0.76} d_{BI}^{-0.63} \nu^{-0.12} \left[ L_B \epsilon_1 \alpha \frac{(1-r_B)}{(1-r_i)} \right]^{0.374} \quad (2)$$

In this equation, the correction factor  $\epsilon_1$  for short tubes can be obtained from Table 1.

Table 1

$L_B/d_{BI}$	1	2	5	10
$\epsilon_1$	1.90	1.70	1.44	1.23

## 2. Conduction Error

Conduction error based on the principles of heat conduction is:

$$\sigma_c = \frac{T_g - T_B}{ch(mL)} \quad (3)$$

$$ch(mL) = \frac{1}{2} (e^{mL} + e^{-mL}) \quad (4)$$

$$m = \sqrt{\frac{\alpha_f U}{\lambda_f}} \quad (5)$$

$$\alpha_f = B_{11} \left( \frac{u \rho_n}{\mu_f g R T_*} \right)^{n_H} d_w (n_H - 1) \quad (6)$$

When the thermocouple wire is parallel to the gas flow,  $n_H = 0.674$  and  $B = 0.0845$ . When it is perpendicular to the gas flow,  $n_H = 0.5$  and  $B = 0.44$ . The thermal conductivity  $\lambda_f$  and viscosity  $\mu_f$  in Equation (6) are determined at a characteristic temperature  $T^*$ .

Equation (3) only applies to a single phase uniform cross-section. A thermocouple, however, is formed by two different materials. In this case, it can be treated as two "thermal resistances" in series. If the average thermal conductivity is adopted:

$$\lambda_w = \frac{\lambda_{w1} + \lambda_{w2}}{2} \quad (7)$$

Equation (3) can still be used to estimate conduction error. If a more rigorous calculation is carried out for parallel soldered thermocouples, the following dual axial heat conduction equation can be used:

$$\sigma_c = \frac{T_g - T_B}{\cosh m_1 L + \left\{ \frac{\lambda_2 m_2 [(\sinh m_2 L - \sinh m_1 L) - (e^{-m_1 L} - e^{-m_2 L}) + 4\alpha_1 \sinh m_2 L]}{\lambda_1 m_1 \sinh m_2 L + \lambda_2 m_2 \sinh m_1 L} \right\} \sinh m_1 L} \quad (8)$$

$$m_1 = \sqrt{\frac{4\alpha_1}{\lambda_{w1} d_w}} \quad (9)$$

$$m_2 = \sqrt{\frac{4\alpha_2}{\lambda_{w2} d_w}} \quad (10)$$

For the same thermocouple wire material, increasing flow velocity and lengthening the immersion depth are the means to reduce conduction error.

### 3. Radiation Error

The radiation error of the thermocouple is:

$$\sigma_r = \frac{eC_o}{\alpha_1} \left[ \left( \frac{T_g}{100} \right)^4 - \left( \frac{T_B}{100} \right)^4 \right] \quad (11)$$

$$\sigma_r = \frac{K_r}{M_P} \left( \frac{T_g}{100} \right)^{-0.18} \left[ \left( \frac{T_g}{100} \right)^4 - \left( \frac{T_B}{100} \right)^4 \right] \quad (12)$$

The  $\alpha_f$  is still calculated using Equation (6) with characteristic length  $d_j$ .

In order to minimize radiation error, it is important to install a sheath and to increase the velocity at the measuring point.

It is necessary to point out that the sheath cannot alter the characteristics of the junction. It only provides a partially suitable environment at the measuring point. It is imperative to open an exhaust hole for such partial environment. Without this exhaust hole, gas inside the sheath can no longer flow. The gas mass around the junction will not circulate. The temperature at the measuring point will be approaching that of the sheath due to heat conduction. In this case the sheath not only does not minimize error but also worsens the situation.

The internal gas velocity cannot be zero. How much should it be? A smaller internal velocity should be used to reduce velocity error, however, a larger internal velocity will minimize radiation and conduction errors. Under such contradictory conditions, the selection of the optimum internal flow velocity becomes the major task in obtaining minimum thermocouple error. The opening of a proper exhaust hole to provide the optimum internal gas flow velocity is practically the core of the design which is also the essence of this paper.

### II. THE OPTIMUM INTERNAL FLOW VELOCITY AND THE DIAMETER OF THE EXHAUST HOLE

The optimum internal flow velocity is the velocity at which minimum thermocouple error is achieved. Thermocouple error is

$$\sigma_T = \sigma_w + \sigma_o + \sigma_R$$

(13)

The velocity error, conduction error and radiation error are as expressed by Equations (1), (3), (6), and (11). They are functions of the internal flow velocity  $u$ . The partial differentiation of  $\sigma_T$  with respect to  $u$  is  $\frac{\partial \sigma_T}{\partial u}$  and let it be zero.

$$\frac{\partial \sigma_T}{\partial u} = \frac{\partial \sigma_w}{\partial u} + \frac{\partial \sigma_r}{\partial u} + \frac{\partial \sigma_R}{\partial u} = 0$$

The corresponding internal flow velocity is the optimum one. Since conduction error is a hyperbolic function, it is difficult to obtain a solution using the above equation. Therefore, the manual calculation of the optimum internal flow velocity is approximated by determining the minimum of  $\sigma_t$  where

$$\sigma_t = \sigma_w + \sigma_R \quad (14)$$

Let  $\frac{\partial \sigma_t}{\partial u} = 0$ , it is possible to obtain the optimum internal flow velocity

$$u_{*} = \left\{ \frac{n_p e C_0 d_{(1-n_p)} v_f^{n_p} \cdot g \cdot C_r \left[ \left( \frac{T_f}{100} \right)^4 - \left( \frac{T_B}{100} \right)^4 \right]}{B \lambda_f (1-r) A} \right\}^{\frac{1}{n_p + 2}} \quad (15)$$

$$T_B = \frac{T_{B0} + T_{Bf}}{2} \quad (16)$$

$$T_{B0} = T^* - \sigma_{wB0} - \sigma_{RB0} \quad (17)$$

$$T_{Bf} = T^* - \sigma_{wBf} \quad (18)$$

In these equations,  $v_f$  and  $\lambda_f$  are determined at a characteristic temperature  $T^*$  while  $C_p$  is obtained at  $T_u$ . During the first round of estimation before the internal velocity is known, it is reasonable to approximate

$$\begin{aligned} T_v &= T \\ T_s &= T_{s0} \end{aligned} \quad (19)$$

The optimum internal velocity must also include conduction error into consideration. Therefore, the calculated internal flow velocity based on Equation (15) is most probably not the real optimum value. If conduction error is large then the calculated internal flow velocity (using Equation (15)) will deviate significantly from the optimum situation. It is possible to correct the problem by increasing the immersion depth of naked wire as well as by properly enlarging the diameter of the exhaust hole to adjust the internal flow velocity.

When gas flows into the sheath, especially at the exhaust hole, significant loss occurs. Therefore, in order to maintain an optimum internal flow velocity  $u_j$  at the junction, the ratio of the cross-section areas of the exhaust hole and the measuring function should be:

$$\frac{n_B f_D}{F_{Bj}} = \frac{M_a}{\zeta_G M} \left[ \frac{1 + \frac{K-1}{2} M^2}{1 + \frac{K-1}{2} M_a^2} \right]^{\frac{K+1}{2(K-1)}} \quad (20)$$

or

$$\frac{n_B f_D}{F_{Bj}} = \frac{q(\lambda_a)}{\zeta_G q(\lambda)} \quad (21)$$

In the above equations, the sheath flux coefficient  $\zeta_G$  depends on the structure, dimension, medium, and the flow condition used in the measurement. Since  $M_G$  is generally less than 0.3 and  $(1-r)$  and  $N_H$  are also less than 1, it is then sufficient to provide the necessary accuracy for total temperature thermocouple probes by approximating  $\zeta_G = 1$ . In the design procedure of total temperature thermocouple probes, the flow loss is frequently neglected to simplify the calculation.

We obtain:

$$\frac{n_B f_D}{F_{Bj}} = \frac{M_a}{M} \left[ \frac{1 + \frac{K-1}{2} M^2}{1 + \frac{K-1}{2} M_a^2} \right]^{\frac{K+1}{2(K-1)}} \quad (22)$$

$$\frac{n_B f_D}{F_{Bj}} = \frac{q(\lambda_a)}{q(\lambda)} \quad (23)$$

Severe blockage problem will occur when the size of the sheath is too large. On the other hand, a small sheath is not only too weak in strength but also less effective due to the heat transfer between the sheath and the junction point. Sometimes such an error is larger than that of an exposed thermocouple. In the design of the sheath, the size of the sheath should be kept to the smallest degree possible without endangering the accuracy of the thermocouple. Usually the distance between the sheath and the junction should be at least 1.5 mm. When the inner diameter (which is the cross-section area of the junction) is selected, the diameter of the exhaust hole to realize the optimum internal flow velocity is

$$d_{Dn} = \sqrt{\frac{4F_{Bj} M_a}{\pi n_B M} \left[ \frac{1 + \frac{K-1}{2} M^2}{1 + \frac{K-1}{2} M_a^2} \right]^{\frac{K+1}{2(K-1)}}} \quad (24)$$

or

$$d_{Dn} = \sqrt{\frac{4F_{Bj}}{n_B \pi} \frac{q(\lambda_a)}{q(\lambda)}} \quad (25)$$

### III. COMPUTER PROGRAMMING

The manual calculation of the optimum internal flow velocity using Equation (15) in the design of a total temperature thermocouple probe sheath not only is lengthy but also brings in error due to the various assumptions made in the simplifica-

tion steps. With the development of electronic computers, these shortcomings are avoided. The following section introduces a computer program written in ALGOL 60 (DJS 14) language to design the best total temperature thermocouple probes (see Appendix 2).

The program first reduces the internal flow velocity from the incoming flow velocity  $\omega$  in 1 m/sec steps to the minimum internal flow velocity  $u_{\min}$ . It calculates the corresponding velocity error, conduction error, radiation error and thermocouple error for each internal flow velocity and compares these errors with those stored in the ERT unit. The variables UG, MG, EWG, ECG, ERG and ETG always store the corresponding values of the internal flow velocity, Mach number in the sheath, velocity error, conduction error, radiation error, and thermocouple error for the smallest thermocouple error encountered, respectively. The optimum internal flow velocity and its corresponding error can then be obtained.

Based on the optimum internal velocity, its corresponding exhaust hole diameter can be chosen. Usually, the nominal diameter of the exhaust hole is determined by the specification of drill bits. Frequently the diameter of the drill bit is slightly larger than the optimum value. This practice in a way compensated the error brought about due to the motion of the sheath. In the situation that a fixed exhaust hole diameter has been assigned, this assigned value is used in all the calculations.

Finally, internal flow velocity and its corresponding error and complex thermal coefficient can be determined based on the diameter of the exhaust hole.

The immersion depth of the naked thermocouple wire can be

chosen by evaluating the results obtained from five different values.

The use of computers to determine the optimum internal flow velocity allocated the difficulties encountered during manual calculation. It also takes conduction error, radiation error and velocity error simultaneously into consideration in determining the optimum internal flow velocity. Due to the fact that calculations are made using pre-chosen internal velocity values, the approximations frequently adopted in the manual processing procedure can be completely avoided.

#### IV. DESIGN AND CALIBRATION

In order to evaluate the optimum design procedure of total temperature thermocouple probes, we have carried out an actual design and calibration practice of a total temperature thermocouple probe to measure the temperature of the exhaust gas of a turbojet engine at temperature  $T_4^*$ .

It has been known that: the static temperature of the combustion gas  $T_4 = 865K$ , velocity  $v = 288 \text{ m/sec}$  ( $M = 0.5$ ), total pressure  $P_4^* = 2 \text{ kg/cm}^2$  and the wall temperature of the diffusion chamber  $T_w = 700 \text{ K}$ . The thermocouple is made using 0.5 mm diameter Ni-Cr and Ni-Al wires connected in parallel with junction diameter  $d_j = 1 \text{ mm}$  and absorption coefficient  $\epsilon = 0.8$ . Immersion depth of exposed wire is selected to be 5 mm. The sheath material used is 1Cr18Ni9Ti with 3 mm inner diameter and 4 mm outer diameter. The thermocouple wire and sheath are parallel to the gas flow and their complex thermal coefficients are  $r = 0.86$ ,  $B = 0.0845$ , and  $n_H = 0.674$ .

The design procedures obtained using a DJS14 computer according to the program listed in Appendix II under conditions

specified above are shown in Tables 2 and 3.

If manual calculation has to be relied upon to obtain the optimum internal flow velocity using Equation (15), a second series of calculations is recommended. This is because that the initial calculation uses the approximations that  $T_j = T_4^*$ ,  $T_u = T_4$ ,  $P_u = P_4$ , and  $T_B = T - \sigma_{WBO}$ . From the static temperature  $T_u = T_4$  it was found that  $C_p = 0.2743 \text{ K cal/kg}^{\circ}\text{C}$ . Based on  $T_4^* = 900.8 \text{ K}$ , it was then found that  $u_f = 3.9478 \times 10^{-6} \text{ kg sec/m}^2$  and  $\lambda_f = 5.4394 \times 10^{-2} \text{ K cal/m hour}^{\circ}\text{C}$ . Plugging all these values into Equation (15), the first order estimation of the value of the optimum internal flow velocity  $U_{G1} = 128.57 \text{ m/sec}$ . From  $U_{G1}$  it is possible to determine the velocity error of the sheath and the junction as well as the static pressure  $P_u$  inside the sheath. Equations (16), (17) and (18) can then be used to obtain  $T_B$ . Using the approximation that  $T_j = T^* - \sigma_w$  and plugging back into Equation (15), a second estimation shows that the optimum internal velocity  $U_{G2} = 95.6 \text{ m/sec}$ . If two exhaust holes are used, it is possible to derive from Equation (24) or (25) that the optimum diameter of the exhaust  $d_{DG} = 1.2 \text{ mm}$ .

The nominal value of the diameter of the exhaust hole depends on the available standard drill bit diameter. When different diameters are selected, the corresponding errors are shown in Table 2.

Table 2

$d_p$ mm	1.2	1.3	1.4	1.5	1.6	1.7	
Calculation Method	Computer	Computer	Computer	Computer	Computer	Computer	
$\sigma_w$ °C	0.525	0.730	0.995	0.98	1.333	1.761	2.304
$\sigma_c$ °C	1.099	0.925	0.780	0.84	0.658	0.554	0.466
$\sigma_r$ °C	1.548	1.399	1.267	1.25	1.148	1.040	0.940
$\sigma_t$ °C	3.172	3.054	3.042	3.07	3.139	3.355	3.710
$\sigma_t$ °C	2.073	2.129	2.262	2.23	2.481	2.801	3.244

When  $d_D = 1.2$  mm,  $\sigma_t$  has a minimum value. However, as the conduction error is being considered,  $d_D$  should be 1.4 mm to achieve the minimum error  $\sigma_{\text{couple}}$  for the thermocouple. Table 3 shows the various errors for different immersion depths with a 1.4 mm diameter exhaust hole. With each 1 mm additional immersion depth, the conduction error decreases by 50%. It is extremely effective to reduce conduction error by increasing the immersion depth.

Table 3

L mm	4	5	6	7	8
$\sigma_w$ °C	0.995	0.995	0.995	0.995	0.995
$\sigma_C$ °C	1.460	0.780	0.417	0.224	0.121
$\sigma_R$ °C	1.179	1.267	1.314	1.139	1.352
$\sigma_T$ °C	3.634	3.042	2.727	2.558	2.468

The variations of  $\sigma_T$  and  $\sigma_t$  as a function of  $d_D$  are shown in Figure 1. It can be observed that near the minimum error region the variation of the thermocouple error with the diameter of the exhaust  $\frac{\partial \sigma}{\partial d_D}$  is very small. Therefore, larger tolerance in the actual fabrication of the designed sheath exhaust hole can be allowed.

In order to provide sufficient contact between the gas flow and the thermocouple wire to minimize conduction error, the exhaust hole should be located right at the root of the naked wire and it should be placed on the same side as the two thermo-electrodes (see Figure 2).

In order to improve the relevant characteristics of the total temperature thermocouple probe, certain precautionary measures must be taken in the design of the structure of the sheath. The inlet of the sheath and the measuring junction point must be sufficiently far apart to avoid the effects due to the contraction of streamlines and the environment in front of the sheath inlet.

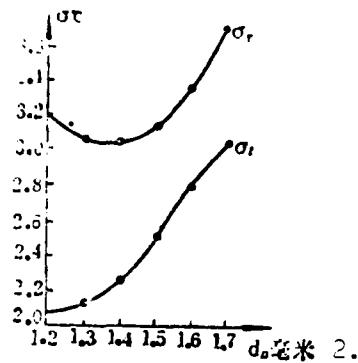


Figure 1.

Key:

2. mm

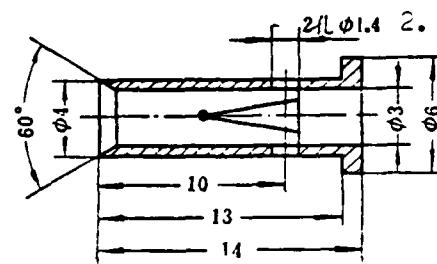


Figure 2.

Key:

2. 2 holes  $\phi$  1.4

The structural design of  $T_4^*$  total temperature thermocouple probe is shown in Figure 2. In the design procedure, the velocity error is  $0.995^\circ\text{C}$  with a corresponding complex thermal coefficient at 0.972.

The  $T_4^*$  total temperature thermocouple probe has been calibrated for its complex thermal coefficient in a wind tunnel. When  $M = 0.5034$ , the potential difference measured using a 355 potentiometer is:

$$E(t^*, 0) = 1.8850 \text{ mV} (46.64^\circ\text{C})$$

$$\Delta E_{\text{mutually calibrated}} = -0.0070 \text{ mV}$$

$$\Delta E_{\text{measured}} = 0.0072 \text{ mV}$$

$$\begin{aligned} E(tg, 0) &= E(t^*, 0) - [\Delta E_{\text{measured}} - \Delta E_{\text{mutually calibrated}}] \\ &= 1.8854 - [0.0072 - 0.007] \\ &= 1.8712 \text{ mV (46.28°C)} \end{aligned}$$

Plugging this value into Equation (1), we get  $r = 0.977$ . The difference between the theoretical and measured values is  $\Delta r = 0.005$ ,  $\Delta r/r = 0.5\%$ .

The  $T_4^*$  total temperature thermocouple probe has also been calibrated with a triple barrier evacuated thermocouple in a hot wind tunnel. Under the condition that  $P^* = 1.0004 \text{ kg/cm}^2$ , static pressure  $P = 0.9639 \text{ kg/cm}^2$  and wall temperature  $T_w = 623 \text{ K}$ , the temperature values obtained with the evacuated thermocouple and the  $T_4^*$  total temperature thermocouple using a potentiometer (Cambridge Model DE) are 877 K and 867.2 K, respectively. Therefore, the error of the  $T_4^*$  total temperature thermocouple  $\sigma_T = 9.8^\circ\text{C}$ . Under the same conditions, the calculated error of the thermocouple  $\sigma_T = 9.3^\circ\text{C}$  which is only  $0.5^\circ\text{C}$  off the calibrated results. It is only 0.06% of the total temperature.

The above presentation indicates that the measured and calculated values of the complex thermal coefficient and thermocouple error are basically identical. This further proves the reliability of this design procedure to a certain degree.

When the axis of the sheath is perpendicular to the direction of gas flow, this situation is far more complicated. There is no practical meaningful calculation method available to date. This paper is only limited to thermocouples located in a sheath with its axis parallel to the gas flow (see Figure 2). Besides,

the present method is more suitable for low or medium flow velocity because under high turbulent flow conditions this method is too conservative in determining the heat transfer error.

#### V. CONCLUSIONS

1. The optimum design of total temperature thermocouple probes is obtained through the determination of the optimum internal flow velocity under which minimum thermocouple error is achieved. The diameter of the exhaust hole is then selected and the design of the sheath structure proceeds. In the calculation of conduction error, the heat conduction characteristics of both thermocouple wires are taken into account by using either a double axial conduction equation or an average thermal conductivity coefficient. In addition, the heat transfer taking place between the inlet and the junction point is limited using the minimum internal flow velocity during the calculation of the optimum internal velocity.

2. The theoretical and measured values of the  $T_4^*$  total temperature thermocouple probe in thermocouple error and complex thermal coefficient are:

$$\sigma_r = 0.5^\circ\text{C}, \frac{\Delta\sigma_r}{\sigma_r} = 0.06\%, \Delta r = 0.005, \frac{\Delta r}{r} = 0.5\%.$$

These results clearly demonstrate the reliability of this design program.

3. This method not only points out the maximum achievable accuracy under a set of conditions, but also can be reduced to actual practice. Thermocouple error is minimized. It does not require prior experience in designing thermocouple probes and it allows larger tolerance in the fabrication of the exhaust hole.

## APPENDIX I

### Symbols

Equation Symbol	Program Symbol	Meaning
$T^*$	TT	total temperature
$T$	T	static temperature
$T_u$	TUTUA	static temperature inside sheath
$T_g$	TG TGA	effective temperature
$T_B$	TWB, TBA, TB	sheath wall temperature
$T_{BO}$	TB $\phi$	outer sheath wall temperature
$T_{BI}$	TBI	inner sheath wall temperature
$T_W$	TW	wall temperature
$T_J$	TJP, TJA, TBB	junction temperature
$P^*$	PT	total pressure
$P$	P	static pressure
$P_u$	PU	static pressure in sheath
$W$	W	gas flow velocity
$M$	M	gas flow Mach number
$a$	A	velocity of sound
$u$	U, UA	internal flow velocity
$u_G$	UG	optimum internal flow velocity
$u_{min}$	UMI	minimum internal flow velocity
$a_u$	A, AUA	speed of sound in sheath
$M_j$	MJ, MJA	Mach number at junction
$M_G$	MG	optimum Mach number in sheath
$K$	K, KBB, KB	insulation factor
$g$	G	gravity
$R$	RG	gas constant
$\mu_f$	MUB, MU, MBB	viscosity
$\lambda_f$	LMF, LMB, LBB	gas thermal conductivity

Equation Symbol	Program Symbol	Meaning
$n_B$	NB	number of exhaust hole
$\sigma_w$	ERW, EW	velocity error
$\sigma_C$	ERC, ARC, ECP	conduction error
$\sigma_R$	ERR, ARR, ERP	radiation error
$\sigma_T$	ERT	thermocouple error
$\sigma_{WBO}$	WB $\phi$	velocity error at outer sheath wall
$r$	R	complex thermal coefficient
$r_j$	RJ	complex thermal coefficient at junction
$r_B$	RB	complex thermal coefficient at barrier
	GAS	type of gas flow
	H	exothermic coefficient
$\gamma_f$		kinetic viscosity
$\lambda_1$	LM 1	thermal conductivity of #1
$\lambda_2$	LM2	thermal conductivity of #3
$\alpha_1$	HL, HLA	exothermic coefficient along the thermocouple wire
$\alpha_j$	HJ, HJA	exothermic coefficient at measuring point
$\alpha_{B0}$	HB $\phi$	exothermic coefficient of barrier outer wall
$\epsilon_1$	EPL, ELB	correction factor for short tubes
$\epsilon_j$	EPJ	absorption coefficient of the measuring point
$\epsilon_B$	EPB	absorption coefficient of the sheath
B	B	equation coefficient
$n_H$	NH	equation exponent
$d_w$	DW	diameter of the thermocouple wire
$d_j$	DJ	diameter of the junction

Equation Symbol	Program Symbol	Meaning
$L_w$	LW, M	immersion depth of the thermocouple wire
	LJ	thermocouple wire length at the cross-section of the junction
$L_B$	LB	inlet to junction distance
$d_{BO}$	DBφ	outer diameter of the sheath
$d_{BI}$	DBI	inner diameter of the sheath
	RLD, RLB	LB/DBI
$d_D$	DD, DDI	diameter of the exhaust hole
$d_{DG}$	DG	optimum diameter of exhaust hole
$d_M$	DM	drill bit diameter
$f_w$	FW	cross-section area of thermocouple wire
$f_D$	FD	area of exhaust hole
$B_J$	FBJ	area of sheath cross-section at junction
$q(\lambda)$		gas dynamic function
$\sigma_{WG}$	EWG	optimum velocity error
$\sigma_{CG}$	ECG	optimum conduction error
$\sigma_{RG}$	ERG	optimum radiation error
$\sigma_{TG}$	ETG	optimum thermocouple error
$\sigma_{WBI}$	WBJ	velocity error at inner sheath wall
$K_r$	KR	radiation correction factor
	SJ	junction type
	SDD	number of immersion depth
	STA	state
	SW	thermocouple state
	SB	sheath state
	N	individual intrusion length numbers

## APPENDIX II

Computer program for the optimum design procedure of total temperature thermocouple probe (omitted)

### REFERENCES

- [1] Haig, L. B., A Design Procedure for Thermocouple Probes. SAE, April 5-8, 1960.
- [2] Moffat, Robert J., Gas Temperature Measurement, Temperature, Its Measurement and control in Science and Industry, Vol. 3, Part 2, 1962.
- [3] Glawe, G. E., Simons, F. S., Stichnev, T. M., Radiation and Several Chromel-Alumel Thermocouple Probe in High-Temperature High-Velocity Gas stream, NACA TN 3766, October 1956.
- [4] Lecture on temperature measurement, Northwest Industrial College [PRC] 1973.
- [5] Tauras, J. A., Some Designs Using sheathed Thermocouple wire for Jet Engine Applications, Temperature, Its Measurement and Control in Science and Industry, Vol. 4, Part 3, 1972, pp. 1805-1810.
- [6] Baas, P. B. R., Mai, K., Trends of Design in Gas Turbine Temperature Sensing Equipment, Temperature, Its Measurement and Control in Science and Industry, Vol. 4, Part 3, 1972, pp. 1811-1821.
- [7] Li Gaode, General methods for analyzing temperature sensor characteristics and calculating for error, unpublished.
- [8] Zhang Suzhen, Liu Zhaoren, Zhu Hong, Techniques for checking pressures and temperatures during testing of engine assemblies and parts, unpublished.
- [9] Li Chunfang, Engine temperature measurenet, unpublished.
- [10] M. A. Mikheyev, translated by Wang Buxuan, Fundamentals of thermal conductivity, Higher education publishing house, 1958.
- [11] Instruction manual for DDC I- Computer.

## Summary

# A Synthesis Technique for Array Antennas of High Directivity and Low Sidelobe

Wan Wei, Huang Jingxi, and Hu Shiming

In modern radar systems an antenna is expected to possess both high directivity and low sidelobe. Known synthesis techniques usually optimize with respect to a single antenna performance index, and good theoretical results have been obtained, but unfortunately such optimization sometimes results in disagreement between theory and practice. In this paper, two kinds of synthesis techniques for array antennas are reviewed with a view to finding a synthesis optimization technique, applicable to the simultaneous consideration of two or more performance indices. These performance indices are: the directivity and/or signal-to-noise ratio (SNR), efficiency, sidelobe level, the positions of null and maximum of sidelobe, etc. This paper differs from past papers in that it considers the optimization with respect to high directivity and low sidelobe simultaneously.

There are two kinds of synthesis technique to achieve high directivity and low sidelobe for array antennas.

The kind of synthesis technique is to apply the matrix theory for array antennas to obtain antenna indices such as directivity  $D$  as follows:

$$D = \frac{|\mathbf{A}\mathbf{l}|}{|\mathbf{B}\mathbf{l}|},$$

where

$\mathbf{l}$ —N-element column matrix for a series of complicated unit excitation functions.

$\mathbf{l}^*$ —conjugate transpose of matrix  $\mathbf{l}$ .

$\mathbf{A}$   $\mathbf{B}$ —Hermitian matrices.

Optimization requires that the maximum of  $D$  and the  $\mathbf{l}$  matrix corresponding to the maximum of  $D$  are to be found. If the maximum directivity is to be obtained for a given sidelobe level, then the constraining relationship to be satisfied for finding the constrained maximum directivity is

$$\frac{E(\theta_0, \phi_0)}{E(\theta_0, \phi_0)} = \frac{|\mathbf{A}\mathbf{l}|}{|\mathbf{B}\mathbf{l}|} = S_0$$

where the subscripts "i" and "o" denote the directions of sidelobe and main-lobe respectively. Optimization under given constraint is accomplished by using Lagrange multipliers.

Applying the matrix theory to obtain optimization is an iterative process, in which adjustments of current or spacing or both are made successively on a given original array, until further adjustments are no longer worthwhile. Past papers<sup>[13]</sup> point out that careful attention should be paid to the convergence of the iterative process, and to the adequacy of the computer's memory capacity and computing speed; but thus far, we have not seen actual numerical application of this technique to optimize with respect to both high directivity and low sidelobe simultaneously.

The second synthesis technique is the iterative sampling method, which can be conveniently used to control shaping-beam radiation pattern, sidelobe behavior, etc. A unique feature of this method is its convergence capability for a solution satisfying a given set of specifications, i. e., the iteration process stops when these specifications are satisfied. Thus it avoids undesirably complicated current distribution on array antennas or excessive spreading out of the beam width.

The basic idea of the sampling iterative method is as follows:

On a given initial pattern  $E^*(\theta, \phi)$  is added, through sampling, a correction pattern  $\Phi(\theta, \phi)$ . The resulting pattern obtained is then the first iterative pattern. If it does not satisfy the requirements of the expectation function  $E^*(\theta, \phi)$ , then the sampling iterative operation is continued further. For the i-th iteration, the total correction pattern is the sum of all the  $2N$  products obtained by multiplying each correction pattern by its correction coefficient:

$$\Delta E^*(\theta, \phi) = \sum_{n=-N}^N a_n^* \Phi[(\theta - \theta_n^*), (\phi - \phi_n^*)],$$

where  $a_n^*$ —the correction coefficient.

$(\theta_n^*, \phi_n^*)$ —the correction pattern center (i. e., sampling point).

After the  $s$ -th iteration, the resulting pattern is the sum of the initial pattern and the  $s$  total correction patterns:

$$E^*(\theta, \phi) = E^*(\theta, \phi) + \sum_{i=1}^s \Delta E^i(\theta, \phi).$$

With more and more iterations,  $E^*(\theta, \phi)$  approaches  $E^*(\theta, \phi)$ .

By a similar derivation, the aperture distribution of the planar array is obtained as follows:

$$I^*(p) = I^*(p) + \sum_{i=1}^s \Delta I^i(p),$$

where

$p$ —the normalization coordinates of the planar array.

$I^*(p)$ —the initial aperture distribution.

$\Delta I^i(p)$ —the correction function of the aperture distribution.

In order to control sidelobe level, the sampling point is taken at the position corresponding to the maximum of the sidelobes. The sampling iterative method may be conveniently programmed for calculating with computers. A numerical example for the sampling iterative method is given in this paper.

So far, the authors have found that only the sampling iterative method is practicable for optimization with respect to both high directivity and low sidelobe simultaneously.

A Synthesis Technique of Array Antennas  
of High Directivity and Low Sidelobe

by

Wan Wei, Huanq Jingxi, and Hu Shiming

Abstract

It is expected in the PDR system that the antenna has high directivity and low sidelobe characteristics. In practice, it is easier to accomplish this goal using spaced arrays than a continuous aperture structure. The synthesis of array antennas can be done in two ways. One is to optimize directivity using theoretical matrix treatment under a sidelobe constraint. The other is a sampling iterative method. With respect to a given initial pattern  $E^0(\theta, \phi)$ , through sampling of the peak value of the sidelobe, a correction function  $\Phi(\theta, \phi)$  is gradually being iterated to approach the expectation function  $E^1(\theta, \phi)$ . The apparent advantages of this method are the convergence capability and great flexibility in a certain application. This paper also includes an actual example of the iterated sampling method.

I. INTRODUCTION

Due to the invention of low noise amplifier and its application in radio-astronomy, communication, and radar, the antennas become the main source of noise for the entire system. Furthermore, typical noise temperature is several times larger than that of the amplifier. This troublesome problem together with the long existing interference problem due to ground objects desperately demand the lowering of the sidelobe level. The synthesis of antennas may also proceed with the optimum SNR value.

Being a combat-oriented radar such as the PDR system, it is frequently required to distinguish radar targets from a very dense background in order to complete searching and begin tracking. This requirement generated lots of interests in high resolution and high directivity antennas. Resolution and directivity are described by the pattern of the radar. High resolution generally indicates a narrow main lobe and low sidelobes. High directivity means that relative to the total radiated energy the mean beam's radiation energy pointing upward is high.

In practice, the synthesis of array antennas is far more flexible than using a continuous antenna structure. In some cases, it is easier to materialize. Earlier work in array antenna synthesis often began with one performance index. Therefore, there are two major areas with an assortment of literature available. These are:

#### Optimum Directivity Synthesis

This was originally pointed out by Schejunooff [1] that theoretically high directivity can always be obtained with arrays of finite dimensions. Later other researchers confirmed that this statement is sound based on theory. Soon after Uzkov [2] attempted to solve the optimization of directivity using a linear transformation technique. Block et al [3] used a differentiation method to approach this same problem and obtained an expression for the maximum directivity of any spaced array. There were other authors reported their results of the effects of various factors on the optimum directivity. For example, Gilbert and co-authors [4] studied the optimum design of antenna directivity as a function of random variations. They considered factors such as sensitivity and common difference. Y.T. Lo et al [5] studied the optimization of directivity and SNR under Q factor constraints

with apparent solutions. Butcher and Unz [6] moved a step forward and investigated the optimization of beam efficiency of non-uniformly spaced arrays. More recently, D. K. Cheng [7,8] systematically described methods for optimization of directivity and SNR. In the synthesis of the array with optimized directivity as mentioned above, the given radiation pattern has not been provided. It is usually found that the solution to the array matrix associated with the optimal directivity has a relatively high sidelobe level in its pattern. This consequence is seldom mentioned. However, such high sidelobes will introduce undesirable interference in the identification of targets.

In the synthesis of beam shape, the control of beam shape is the first objective. Usually, such a control is obtained at the expense of improved directivity. Dolph [9] developed the first optimization concept in this synthesis technique; for a fixed band with a minimum sidelobe level can be attained or vice versa. Thereafter, Stegen [10] and Maas [11] actually worked out the optimization of array antennas with a continuous aperture distribution. The most remarkable results were reported by Taylor [12] who developed Dolph's technique another step forward and used it in the synthesis of array with a continuous aperture distribution. He began to attack this problem with a functional approach using an entire function to approximate the ideal spatial factor [13]. The results obtained are far more superior than those derived from Dolph's technique and attracted a lot of attention. Taylor's results are not only applicable to linear antennas but also can be extrapolated to circular aperture antennas. Recently Rhodes [14] studied the Taylor distribution and adopted Taylor's suggestion to make  $\alpha = 1$  in order to maintain 0 at the edge of the aperture and keeping the characteristics of  $\alpha = 0$ . Goto [15] used the Gegenbauer polynomial to synthesize array antennas for high selectivity.

and low sidelobe. The selected radiation filled is described using a Gegenbauer polynomial and the aperature distribution is a Jacobi polynomial. Equations suitable for computer calculation are derived.

Our goal is to combine the high directivity and low sidelobe level during the synthesis in order to meet the practical requirements. Drane et al<sup>(16)</sup> have described the process for the simultaneous optimization of the array antennas and control of the origin of the directional pattern. Their method placed a lot of emphasis on the combination of the optimization and the control of the radiation pattern without considering the sidelobe level. From past experience, people believe that the solution to the above problem can be obtained using matrix theory of the array antennas. Low sidelobe level is introduced as a parametric constraint in the optimization of directivity and SNR during the synthesis. Lagrange multiplier is used to complete the calculation. In practice the more given constraints are present, the more complicated the calculation of the solution becomes using Lagrange multiplier. Under given constraints, the decrease of the order of the matrix is equal to the number of constraints. This actually simplifies the computer work involved. For a large array, however, the use of a computer to iterate a solution to such a large matrix causes more concern regarding memory storage space and calculating speed.

Another method is the iterative sampling technique.<sup>(13,19)</sup> The basic idea of this method is to begin with a given initial pattern  $E^o(\theta\phi)$ , through sampling and a correction factor  $\phi(\theta\phi)$ , to obtain the expectation function  $E^d(\theta,\phi)$  after many iterations. It must be noticed that when this method is used to lower the sidelobe of array antennas, convergence during the iteration process is very important. In the meantime, the

choice of the original pattern has a great effect on the iteration steps as well as the results.

These two methods are going to be discussed in detail in Sections §2 and §3, respectively. After a careful comparison, the latter requires far less calculation. It is also much easier to materialize. The flexibility of the synthesis of the shaping-beaming is also greater. This paper will then provide an example of this method at the end.

## 2. The Matrix Synthesis of Array Antennas

### 1. General Principles

As shown in Figure 1, in three dimensional space with  $N$  identical elements arbitrarily scattered, the  $n$ th element at coordinates  $(r_n, \theta_n, \phi_n)$  has an expectation value which is expressed as  $i_n e^{j\psi_n}$ . Based on this expression, the radiation field generated by this array of  $N$  elements can be written as:

$$E(\theta, \varphi) = \sum_{n=1}^N i_n e^{j(\psi_n + kr_n \cos \alpha_n)} \quad (1)$$

where  $k = \frac{2\pi}{\lambda}$ ,

$$\cos \alpha_n = \sin \theta \sin \theta_n \cos(\varphi - \varphi_n) + \cos \theta \cos \theta_n \quad (2)$$

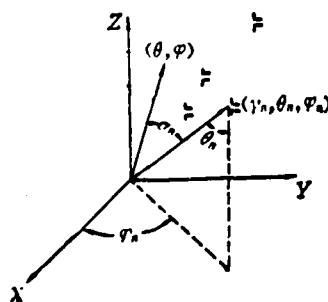


Figure 1. Reference Coordinate System of an Arbitrary Array.

The definition of directivity of an antenna is:

$$D = \frac{\text{Max. Radiation Power Density of the Mainlobe in Direction}}{\text{Average radiation power density}} \quad (3)$$

From (1) we know the radiation power density in direction  $(\theta_0, \phi_0)$  :

$$P(\theta_0, \phi_0) = |E(\theta_0, \phi_0)|^2 = \sum_{n=1}^N \sum_{m=1}^N I_n I_m^* e^{-jk(r_m \cos \alpha_{0m} - r_n \cos \alpha_{0n})}$$

Let's introduce two  $N$  dimensional column vectors  $\mathbf{I}$  and  $\mathbf{a}_0$ .  $\mathbf{I}$  corresponds to unit excitation function i.e.

$$\mathbf{I} = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix}, \quad I_1 = i_1 e^{j\varphi_1}, \quad I_2 = i_2 e^{j\varphi_2}, \quad \dots I_N = i_N e^{j\varphi_N}$$

$\mathbf{a}_0$  is a column matrix formed by the phase factor of the element structure due to spacing between elements:

$$\mathbf{a}_0 = \begin{bmatrix} e^{-jkr_1 \cos \alpha_{01}} \\ e^{-jkr_2 \cos \alpha_{02}} \\ \vdots \\ e^{-jkr_N \cos \alpha_{0N}} \end{bmatrix}$$

where  $\cos \alpha_{0n} (n = 1, 2, \dots, N)$  is obtained from (2) with  $\theta = \theta_0$ , and  $\varphi = \varphi_0$ . A  $N \times N$  square matrix  $\mathbf{A}$  is defined as

$$\mathbf{A} = \mathbf{a}_0 \cdot \mathbf{a}_0^*$$

where "+" is the complex conjugate.  $\mathbf{A}$  is the product of two column vectors and its element is

$$A_{mn} = e^{-jk(r_m \cos \alpha_{0m} - r_n \cos \alpha_{0n})}$$

$A$  is Hermitian i.e.  $A_{mn} = A_{nm}^*$

Therefore the radiation power density in the  $(\theta_0, \phi_0)$  direction can be written as:

$$P(\theta_0, \phi_0) = \mathbf{I}^* \mathbf{A} \mathbf{I}$$

The average radiation power density of the antenna in the radiation pattern is

$$P_r = \frac{1}{4\pi} \int_0^{2\pi} d\phi \int_0^{\pi} |E(\theta, \phi)|^2 g(\theta, \phi) \sin \theta d\theta$$

where  $g(\theta, \phi)$  is the single element power direction pattern function and  $g(\theta_0, \phi_0) = 1$ . Let's define another  $N \times N$  square matrix  $\mathbf{B}$ , with its element.

$$B_{mn} = \frac{1}{4\pi} \int_0^{2\pi} d\phi \int_0^{\pi} g(\theta, \phi) e^{-jk(r_m \cos \alpha_m - r_n \cos \alpha_n)} \sin \theta d\theta$$

This square matrix is symmetrical and Hermitian. The average radiation power density is then simple to express as

$$P_r = \mathbf{I}^* \mathbf{B} \mathbf{I}$$

Henceforth, the directivity of an antenna (3) can be re-written as

$$D = \frac{P(\theta_0, \phi_0)}{P_r} = \frac{\mathbf{I}^* \mathbf{A} \mathbf{I}}{\mathbf{I}^* \mathbf{B} \mathbf{I}} \quad (4)$$

which is the ratio of two Hermitian matrices. From (4) we can conclude that with fixed geometry, dimension, working wavelength, and scanning angle, the optimization effort becomes the search for a special  $\mathbf{I}$  under which the improvement of directivity  $D$  is greatest. On the other hand, the purpose of optimization

is to search for the current distribution on the array with maximum D value. There are two advantages using the above expression. The first is that the necessary calculation can be carried out using a computer with existing programs for matrix. Perhaps more importantly, the extreme value of the characteristic index in this equation is easy to determine. The solution can be derived from the orthogonality relation  $\mathbf{A} - \lambda \mathbf{B}$  formed by two real second order type of matrices  $\mathbf{A}$  and  $\mathbf{B}$ . When the corresponding vector  $\mathbf{I}$  satisfies the homogeneous equation:

$$\mathbf{A}\mathbf{I} = \lambda \mathbf{B}\mathbf{I} \quad (5)$$

the characteristic equation is

$$\det(\mathbf{A} - \lambda \mathbf{B}) = 0 \quad (6)$$

and the maximum value can be obtained.

Equation (6) provides solutions to all the extreme values. With respect to the optimization of directivity, however, there is one non-zero value which is

$$D_M = \lambda_M = \mathbf{a}_0^* \mathbf{B}^{-1} \mathbf{a}_0 \quad (7)$$

The corresponding vector  $\mathbf{I}$ , which is the current distribution in the array at maximum directivity, can be written as

$$\mathbf{I} = \mathbf{B}^{-1} \mathbf{a}_0 \quad (8)$$

Equation (7) and (8) provide a pretty rigorous solution to maximum directivity.

## 2. Optimization under constraints

The optimization of one particular index has lots of disadvantages. For example, the synthesis of maximum directivity leads to an array of antennas with very low radiation efficiency. Even between neighboring elements, there is a significant distribution of current with opposite phase factors.

Therefore, it is usually done to sacrifice some directivity in order to improve other indices of the array antennas for better performance of the entire system. Therefore, the problem has been changed to the optimization of directivity under certain parametric constraints.

In order to simplify our discussion on constraint conditions, let us express (1) in terms of the spatial vector  $\mathbf{a}$  and the excitation vector  $\mathbf{I}$ :

$$\mathbf{E}(\theta, \varphi) = \mathbf{I}^* \mathbf{a} \quad (9)$$

When constraints exist in the sidelobe, the ratio between the mainlobe and the sidelobe in the synthesis then becomes fixed:

$$\frac{E(\theta_i, \varphi_i)}{E(\theta_0, \varphi_0)} = \frac{\mathbf{I}^* \mathbf{a}_i}{\mathbf{I}^* \mathbf{a}_0} = S_i \quad i = 1, 2, \dots, M$$

where "i" and "0" represent the sidelobe and the mainlobe, respectively. This equation can also be written as

$$\mathbf{I}^* \mathbf{W}_i = 0 \quad (10)$$

where  $\mathbf{W}_i = \mathbf{a}_i - S_i \mathbf{a}_0$ .

The problem now is to determine the array distribution with maximum directivity  $D_x$  and under constraint condition (10). The usual technique is to use Lagrange multiplier to obtain a solution. A Lagrange Multiplier  $\Lambda$  is multiplied to (10) and the resultant function is added to (4) to form a Lagrange function which is in equilibrium with respect to

the vector  $\mathbf{I}$ .

$$\mathbf{AI} - \mathbf{DBI} + \sum_{i=1}^M \mathbf{A}_i \mathbf{W}_i = 0 \quad (11)$$

or

$$\mathbf{AI} - \mathbf{DBI} + \mathbf{HA} = 0 \quad (11a)$$

Where  $\mathbf{H}$  is a  $N \times M$  matrix with elements  $W_1, W_2, \dots, W_M$ .  
 $\mathbf{A}$  is a vector with elements  $A_1, A_2, \dots, A_M$ .  $\mathbf{HA}$  in equation (11a) is the constraint vector. If  $\mathbf{I}_0^t$  is the complete orthogonal base unit vector set of the unconstraint matrix equation

$$\mathbf{AI}_0^t - \lambda_{0i} \mathbf{BI}_0^t = 0$$

then any vector  $\mathbf{I}$  can be expressed as the linear combination of this complete orthogonal unit vector set

$$\mathbf{I} = \sum_{i=1}^M \alpha_i \mathbf{I}_0^t$$

Plugging this into (11) and multiplying  $(\mathbf{I}_0^t)^*$  on both sides,  $\alpha_i$  becomes the following by orthogonality

$$\alpha_i = \frac{(\mathbf{I}_0^t)^* \mathbf{HA}}{D - \lambda_{0i}}$$

and the solution to  $\mathbf{I}$  vector is

$$\mathbf{I} = \sum_{i=1}^M \frac{(\mathbf{I}_0^t)^* \mathbf{HA}}{D - \lambda_{0i}} \mathbf{I}_0^t \quad (12)$$

where  $\mathbf{A}$  is unknown. Substituting (12) into (10), it is possible to obtain the  $M$  linear equations for  $\mathbf{A}$ .  $\mathbf{A}$  can be obtained by setting the determinant of the coefficients of those linear equations to zero. The determinant is a function of  $D$ . After

some not too complicated transformations, its only solution provides the excitation vector:

$$\mathbf{I} = \frac{1}{K} \sum_{i=1}^N \alpha_i \mathbf{I}_i^c \quad (13)$$

where

$$K = \left[ \sum_{i=1}^N \alpha_i^2 \right]^{1/2}.$$

Constraint system and non-constraint system are inter-related [20]. With respect to directivity, it can be proven that

$$D_m^c \leq D_m \quad (14)$$

The directivity  $D_m^c$  under constraints is always less than the value  $D_m$  obtained without any constraint. For the designer of arrays, they are more interested in minimizing the loss in directivity in order to maintain control in the expectation pattern.

In the synthesis of array antennas using matrix theory, optimization is achieved by repeated iteration. The basis for synthesis is the original array structure. Through the perturbation of a variable which affects the array structure, repeated iteration is carried out until an optimum condition is reached. It is worthwhile noting that the convergence capability during iteration is very important.

### §3. The Synthesis of Array Antennas Using an Iterative Sampling Method

Using an iterative sampling method in the synthesis of array antennas [16-19] is a new technique applied to the synthesis process. This method is very flexible which also

allows the easy control of shaping beam radiation pattern and the suppression of sidelobe level. Stutzman [17] pointed out that the iterative sampling technique always provides convergence capability to any solution for a certain technique. When the requirements of the technique are satisfied, the calculation process is terminated to avoid any over design due to the broadening of the directive pattern of the mainlobe or the complication of the current distribution. The iteration process involves calculations of a series of lower order functions. Compared to the first method, this iterative sampling technique is more suitable for both large and small array applications.

When the iterative sampling method is used to optimize directivity and lower sidelobe of array antennas, it is extremely important to choose an original directive pattern with better quality because it would not only reduce the load of the iteration process but also leads to more satisfactory results.

### 1. Basic Theory of the Iterative Sampling Method

Iterative sampling method is based on the linear stability of the antenna. Therefore, if  $E^d(\theta, \phi)$  is the expectation spatial pattern, then the search for this function can begin with any direction pattern  $E^o(\theta, \phi)$  which is approximately close to  $E^d(\theta, \phi)$ .  $E^o(\theta, \phi)$  is defined as the original directive pattern which is either a known function or an experimental pattern. Iterative sampling is to add a series of correction patterns to  $E^o(\theta, \phi)$ . If the synthesized pattern after one iteration is not meeting the requirements, more iteration steps can be carried out. Henceforth, in the  $i^{th}$  iterations, the corrected directive pattern can be

expressed by the sum of correction coefficients multiplied by their corresponding corrected directive patterns:

$$\Delta E^{(i)}(\theta, \varphi) = \sum_{n=1}^N a_n^{(i)} \phi[(\theta - \theta_n^{(i)}), (\varphi - \varphi_n^{(i)})] \quad (15)$$

where  $a_n^{(i)}$  is the correction coefficient;  $\phi[(\theta - \theta_n^{(i)}), (\varphi - \varphi_n^{(i)})]$  is the corrected pattern with its maximum at  $(\theta_n^{(i)}, \varphi_n^{(i)})$ , and  $N$  is the sampling number. After successive iterations, the correction coefficient  $a_n^{(i)}$  approaches 0 when the requirements are met. The synthesized direction pattern after  $S$  number of iterations is the sum of the original pattern and all those correction patterns, i.e.

$$E^{(S)}(\theta, \varphi) = E^{(0)}(\theta, \varphi) + \sum_{i=1}^S \Delta E^{(i)}(\theta, \varphi) \quad (16)$$

The corresponding aperture distribution can be obtained through Fourier transformations according to

$$\begin{aligned} E^*(\theta, \varphi) &\Leftrightarrow I^*(p) \\ \phi(\theta, \varphi) &\Leftrightarrow I(p) \end{aligned} \quad (17)$$

From the stability of antennas one gets

$$\phi[(\theta - \theta_n^{(i)}), (\varphi - \varphi_n^{(i)})] \Leftrightarrow I(p) e^{jk(x \sin \theta_n^{(i)} \cos \varphi_n^{(i)} + y \sin \theta_n^{(i)} \sin \varphi_n^{(i)})} \quad (18)$$

Where  $p$  is the unified coordinate of antenna aperture, equation (18) indicates that the spatial transformation of the radiation pattern to  $(\theta_n^{(i)}, \varphi_n^{(i)})$  corresponds to the phase shift of a single incidence on the focal plane.

Therefore, equation (15) can be rewritten as

$$\begin{aligned}\Delta I^i(p) &= \sum_{n=-N}^N a_n^{(i)} I^{(i)}(p) \\ &= I(p) \sum_{n=-N}^N a_n^{(i)} e^{j k (x \sin \theta_n^i \cos \varphi_n^i + y \sin \theta_n^i \sin \varphi_n^i)}\end{aligned}\quad (19)$$

Consequently, after S iterations we can write based on (16):

$$I^{(S)}(p) = I^0(p) + \sum_{i=1}^S \Delta I^i(p) \quad (20)$$

The proper choice of  $a_n^{(i)}$  in the above equations should be based on the expectation value in the  $(\theta_n^i, \phi_n^i)$  direction. It should satisfy using the unified expression:

$$\frac{|E^{(i-1)}(\theta_n, \varphi_n) - a_n^{(i)} \phi(\theta_n^i, \varphi_n^i)|^2}{|E^{(i-1)}(\theta_n, \varphi_n) - a_n^{(i)} \phi(\theta_n^i, \varphi_n^i)|^2} = (t - \epsilon)^2 \quad (21)$$

where  $t$  is the expectation value.  $\epsilon$  is a factor indicating precision. As  $\epsilon \ll 1$ , we get

$$a_n^{(i)} = \frac{E^{(i-1)}(\theta_n, \varphi_n) - t E^{(i-1)}(\theta_n, \varphi_n)}{\phi(\theta_n^i, \varphi_n^i) - t \phi(\theta_n^i, \varphi_n^i)} \quad (22)$$

For the suppression of sidelobe level, the condition should be:

$$|E^{(i-1)}(\theta_n, \varphi_n) - a_n^{(i)} \phi[(\theta - \theta_n^i), (\varphi - \varphi_n^i)]| < |E^{(i-1)}(\theta_n, \varphi_n)| \quad (23)$$

The sign of the factor  $a_n^{(i)}$  is thus determined.

The selection of a correction pattern is in principle arbitrary. However, the correction function usually has a simpler expression than that of the original pattern. Under such conditions, it can be obtained from (19) and (20) that

$$I^{(S)}(p) = I^0(p) \left\{ 1 + \sum_{i=1}^S \sum_{n=-N}^N a_n^{(i)} e^{j k (x \sin \theta_n^i \cos \varphi_n^i + y \sin \theta_n^i \sin \varphi_n^i)} \right\}$$

where the term in the parentheses is the factor in the original aperture distribution.

## 2. The Synthesis of Low Sidelobe Antennas Using Iterative Sampling Method.

(1) Consideration of the Original Direction Pattern at Z+ has been pointed out before that the original pattern  $E^o(\theta, \phi)$  when used to obtain the expectation pattern  $E^d(\theta, \phi)$  by an iterative sampling method should approximate the expectation pattern. Therefore, it is better to choose original functions of high quality for the synthesis of high directivity and low sidelobe array antennas in order to minimize the iteration calculations and to obtain satisfactory and meaningful results.

It is well known that an uniform aperture distribution antenna has a high sidelobe level in its distant field. This is a limiting factor in actual practice. Non-uniform aperture distribution antennas, however, were first optimized by Dolph using Yeghweb's polynominal to describe the synthesized distant field. In that expression, all the sidelobes are at the same level. With larger arrays and more elements, the radiation power of the sidelobes occupies a larger portion of the total radiation power which lowers the directivity of the array antennas. Furthermore, as the array element number  $N$  increases the amplitude at the fringe is far greater than that of neighboring elements. This is very difficult to realize in practice. The quasi-optimization procedure developed by Taylor is a compromise between a uniform distribution pattern and the Dolph distribution pattern. This means that the sidelobe next to the mainlobe has the optimized directivity characteristics of those of the Dolph array. The distant sidelobes, on the other hand, have the characteristics of the

uniform distribution arrays. The former makes the mainlobe narrower, while the latter decreases the amplitude at the edge of the array to reduce its effectiveness to the directive pattern. The radiation power of the sidelobes, consequently, becomes a lesser fraction of the total radiation power to increase directivity of the array antennas.

Recently, Goto [15] used Gegenbauer polynominal to synthesize a high directivity low sidelobe array antenna with

$$E(u) = G_{k-1}^t(z_0 \cos u) \quad (24)$$

where  $u = \frac{\pi d}{\lambda} (\sin \theta \sin \theta_0)$ ,  $z_0$  is a parameter regulating the sidelobe level similar to the one used in the Dolph synthesis.  $G_n^t$  is the Gegenbauer polynominal. It is not difficult to find using Fourier transforms that the amplitude distribution of the array element is:

$$I_n = \binom{K-n}{1}_n F(-n, n-K+1, 2-t-K, \frac{1}{z_0^2}) \quad (25)$$

which is a Jacobi polynominal and it can be transformed to

$$I_n = \binom{K-n}{K-n+t-1} \binom{t}{1}_n F(-n, n-K+1, t, a) \quad (25a)$$

where  $a = (z_0^2 - 1)/z_0^2$ . Equation (25) is suitable for numerical calculations. In reality it is not too different from 1; therefore  $a$  is very small. Equation (25) becomes:

$$I_n = \binom{K-n}{K-n+t-1} \left[ \binom{t}{1}_n + (K-n-1) \binom{t+1}{1}_n \sum_{s=0}^{n-1} \binom{K-n-s-1}{2}_n a^{n-1-s} \right] \quad (26)$$

where (26) can be calculated using a computer. When  $t = 0$ , it becomes a Chebyshev distribution. Equation (26) becomes identical to the results obtained by Maas.

As discussed above, as the original pattern of a low sidelobe and a fixed improved directivity array antenna, the results obtained by Taylor or Goto can be selected in the synthesis.

## (2) Example

In the control of sidelobe levels, the iterative sampling point is chosen at the maximum sidelobe level of the original pattern. The main lobes of the iterative correction patterns also coincide at the same point as shown in Figure 2. In this case, the  $t$  value in Equation (2) is the expectation value of the sidelobe level. As an example, a circular aperture array is synthesized with a radius of 350 MM. The sidelobe level must be lower than -35 dB. The original pattern  $E^\circ(\theta\phi)$  has been chosen as the Taylor function with sidelobe level at -25 dB. An uniform radiation pattern is used as the correction pattern  $\Phi(\theta\phi)$ . Iterative sampling is taking place on the plane where  $\theta = c$ . The suppression of sidelobe level is symmetric with respect to the main beam. The distant area pattern can be obtained using iterations by a computer according to equations (15), (16) and (21) - (23). The corresponding aperture distribution can be simplified based on (20) as:

$$I^{(n)}(p) = I^\circ(p) + 2 \sum_{i=1}^n \sum_{j=1}^m \dots \quad (27)$$

where  $U_n^i = k \sin \theta_n^i$ .  $I^\circ(p)$  is a Taylor aperture distribution at -25dB. The final aperture distribution can be obtained using (27). The calculation program is as follows:

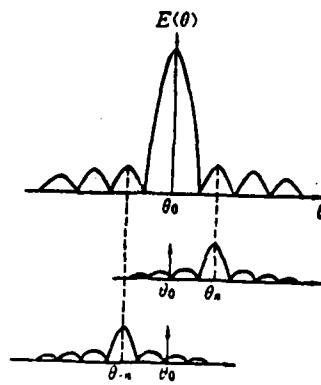
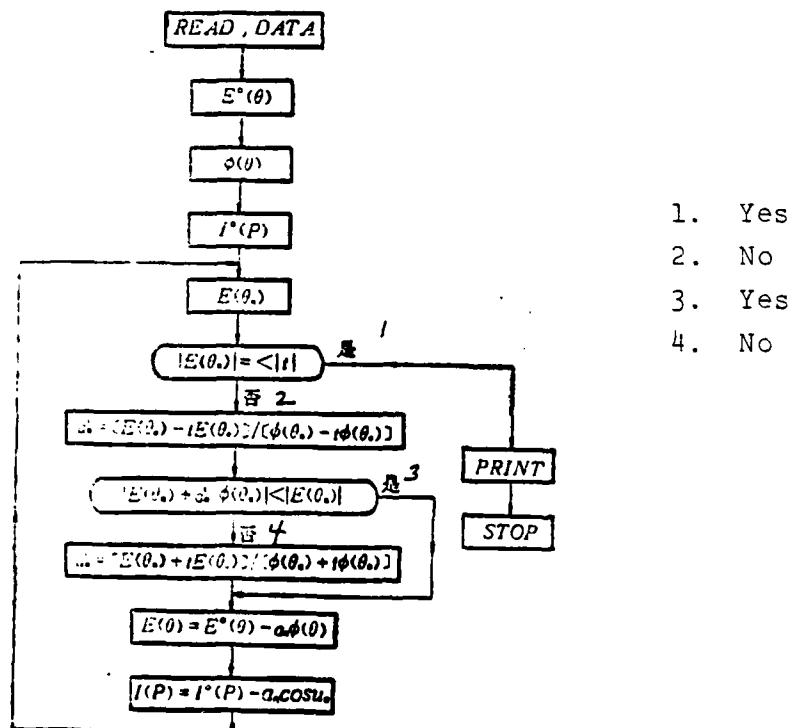


Figure 2. The Iterative Synthesis of Low Side lobe Pattern.



The convergence capability of the given technique must be good. Figure 3 shows the patterns before and after the iterations. The dotted line is the -25dB Taylor distribution and the solid line is the results of the iterative sampling method. The physical meaning of the patterns is obvious.

With respect to the characteristics after iterations, Reference [18] explained the aperture distribution. Reference [19] compared this result with the Dolph distribution. It pointed out that the mainlobe is broader using the iterative sampling method but its directivity is higher than that of the Dolph distribution.

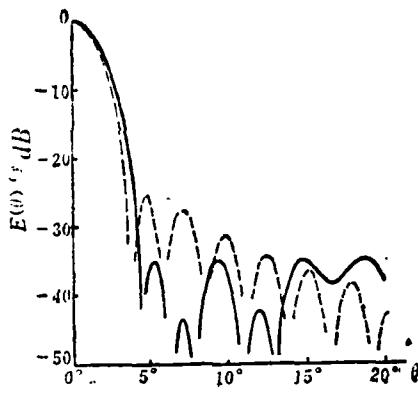


Figure 3. Original Pattern (dotted line) and the Iterative Synthesized Pattern (solid line).

#### 4. CONCLUSIONS

Iterative sampling is in principle the calculation of some elementary functions, while the matrix theory optimization method involves a series of matrix manipulation.

Especially when the array is large, the calculation becomes complicated. In comparison, iterative sampling not only is more flexible but also involves far less calculation. The selection of sampling point in the control of sidelobe level follows a certain rule. This paper reports the results obtained from a DJS-130 computer which coincide with our expectation.

#### REFERENCES

- [1] Scheikunoff, S. A.. A mathematical theory of linear arrays. Bell Syst. Tech. J., Vol. 22, January 1943, pp. 80-107.
- [2] Узков, А. И.. К вопросу об оптимальной направленных антенн. Док. Акад. Наук. СССР., ТОМ. 53, 1946, СТР. 35-38.
- [3] Bloch, A., Medhurst, R G., and Pool, S. D.. A new approach to the design of superdirective aerial arrays. J. IEE, Vol. 100, Pt. II, Sept. 1953, pp. 303-314.
- [4] Gilbert, E. N., and Mergas, S. P.. Optimum design of directive antenna arrays subject to random variations. Bell Syst. Tech. J., Vol. 34, May 1955, pp. 637-663.
- [5] Lo, Y. T., Lee, S. W., and Lee, Q. H.. Optimization of directivity and signal-to-noise ratio of an arbitrary antenna array. Proc. IEEE, Vol. 54, Aug 1966, pp. 1033-1045.
- [6] Butler, J. K., and Uzz, H.. Optimization of beam efficiency and synthesis of nonuniformly spaced arrays. Proc. IEEE, Vol. 54, December 1966, pp. 2007-2008.
- [7] Tseng, F. I., and Cheng, D. K.. Optimum scannable planar arrays with an invariant sidelobe level. Proc. IEEE, Vol. 56, November 1968, pp. 1771-1778.
- [8] Cheng, D. K.. Optimization techniques for antenna arrays. Proc. IEEE, Vol. 59, December 1971, pp. 1664-1674.
- [9] Dolph, C. L.. A current distribution for broadside array which optimizes the relationship between beam width and sidelobe level. Proc. IRE, Vol. 34, June 1946, pp. 335-348.

- [10] Stegen, R. J., Excitation coefficients and beam widths of Tschebyshev array, Proc. IRE, Vol. 41, November 1953, pp. 1671-1674.
- [11] Von der Maas, G. J., A simplified calculation for Dolph-Chebyshev array, J. Appl. Phys., Vol. 25, January 1954, pp. 121-124.
- [12] Taylor, T. T., Design of line source antennas for narrow beamwidth and low sidelobes, IRE Trans. Antennas and Propagat., Vol. AP-3, January 1955, pp. 16-28.
- [13] Taylor, T. T., Design of circular apertures for narrow beamwidth and low sidelobes, IRE Trans. Antennas and Propagat., Vol. AP-8, January 1960, pp. 17-22.
- [14] Rhodes, D. R., On the Taylor distribution, IEEE Trans. Antennas and Propagat., Vol. AP-20, January 1972, pp 143-146.
- [15] Goto, N., A synthesis of array antennas for high directivity and low sidelobes, IEEE Trans. Antennas and Propagat. Vol. AP-16, May 1972, pp. 427-431.
- [16] Hyneman, R. F., A technique for the synthesis of line source antenna patterns having specified sidelobe behaviour, IEEE Trans. Antennas and Propagat., Vol. AP-16, May 1968, pp. 430-435.
- [17] Stutzman, W. L., Synthesis of shaped-beam radiation patterns using iterative sampling method, IEEE Trans. Antennas and Propagat., Vol. AP-19, Januray 1971, pp. 36-41.
- [18] Stutzman, W. L., Sidelobe control of antenna patterns, IEEE Trans. Antennas and Propogat., Vol AP-20, January 1972, pp. 102-104.
- [19] Сафонов, Д.А., и Школьников, А.М., Синтез амплитудно-фазовых распределений в произвольных решетках дающий разномерное приближение к заданной диаграмме направленности, РИЭ., ТОМ 19. № 1, 1971, СТР. 10.
- [20] F.R. Gandemaxe, translated by Ke Zhao, Matrix theory, Vol 1, Higher education publishing House, (1955), pp. 315-324.

## Summary

# Model Method of State Estimation

*Dai Guanzhong*

The problem of estimating the state of a stochastic dynamical system from noisemixed observed values taken from the state is of central importance in modern control theory. The purpose of this paper is to present, in accordance with the relationships of correspondence between the dynamical systems (original) and the filters (model), an unified and direct approach to derive the recursive equations for discrete-time and continuous-time, linear and non-linear, unbiased and minimum-variance filters. For the nonlinear systems, only the second-order approximation suboptimal filters are studied in this paper.

The reasoning followed by this paper is as follows. A linear or non-linear dynamical system, the state of which is to be estimated, is named the original system and characterized by difference equations (the discrete-time system) or differential equations (the continuous-time system), therefore the linear or nonlinear estimator (filter), named the model system, is likewise a dynamical system characterized by the corresponding difference or differential equations.

According to this reasoning, it is obvious that the structure of the linear or nonlinear filter is the same as the structure of the linear or nonlinear dynamical system.

The next step is to determine the parameters of the linear or nonlinear filter.

In the case of linear discrete-time or continuous-time estimation problem, the parameters can be determined by the performance criteria of state estimation: unbiasedness and minimum-variance. First, by the unbiasedness requirement, the filter should have the same number of dimensions as that of the dynamical system and thus form an unbiased filter. Secondly, by the minimum-variance requirement, we can determine the optimal gain matrix and obtain an unbiased and minimum-variance optimal filter. Thereupon we obtain the well-known Kalman equations for the discrete-time filter, or the Kalman-Bucy equations for the continuous-time filter.

Although the above results for the linear filter are not new, the model method, however, is of help in the design of suboptimal nonlinear filter, and

the procedure is much the same. First, by the unbiasedness requirement, we can determine the compensating terms for biasedness and form an approximate unbiased filter. If the first-order approximation is chosen, the compensating terms vanish and the case of extended Kalman filter obtains. In order to improve upon the accuracy of the nonlinear filter, the second-order approximation unbiased filter is chosen. Secondly, by the minimum-variance requirement, we can determine the suboptimal gain matrix and obtain a second-order approximation unbiased and minimum-variance suboptimal filter.

Model Method of State Estimation

by

Dai Guanzhong

ABSTRACT

In accordance with the relationships between the stochastic dynamical system (original) and the corresponding estimation value or filters (model), it is possible to derive the recursive equations for discrete-time and continuous-time, linear and non-linear, unbiased and minimum-variance filters using an unified and direct approach. For the nonlinear systems, only the second order approximation suboptimal filters are studied.

I. INTRODUCTION

The estimation of the state is one of the central topics in modern control theory. An optimal practical meaningful control system very often is the feedback type. The best control signal given by a computer to the controlled system is a function of the state of that system. However, due to the restrictions in practical engineering, the present monitoring devices can only provide a portion of the information of the state. Furthermore, in the monitoring process it is almost unavoidable to pick up accompanying stochastic noises. State estimation is the process of information of the state of a system from insufficient and noise-mixed observed values taken from the state. Mathematically speaking, this belongs to a statistical data processing problem.

Methods frequently used in state estimation are the Gauss least square method, the Wiener filtering method, and the Kalman filtering method. These three are inter-related somehow. From the angle of data processing, there are two techniques viz. the classical "batch processing method" and the modern "recursion processing method". The latter happens to be the process used in the Kalman filtering method [1,2]. The advantage of the recursion method is that it saves computer memory space and reduces actual calculation time. It is more suitable for practical applications.

The conceptual basis of the recursion method lies in the relationships of correspondence between the dynamical system and the filters. In other words, the state of the dynamical system can be described by differential (the continuous-time systems) or difference (the discrete-time systems) equations. Therefore, the estimator (the model system) is also characterized by the corresponding differential or difference equations.

In this paper the filter is considered as a model of the dynamic system. Direct engineering methods are used to derive equations for unbiased, minimum-variance and most meaningful filters. For a linear system, it is known as the Kalman equation (the discrete-time system) or the Kalman-Bucy equation (the continuous time equation). Although there is no new result here, yet this paper is of help in the design of suboptimal nonlinear filters. References [3,4] reported similar studies; however, results are incomplete.

The second section of this paper discusses the filter model for a linear discrete-time system which led to the Kalman filter equation. The third section is a study of

the filter model of nonlinear discrete-time systems which leads to the derivation of a second order approximation suboptimal filter equation for nonlinear discrete-time systems. The fourth section is a study of the filter model of the linear continuous-time system to obtain the Kalman-Bucy filter equation. Finally, section six is a study of the nonlinear continuous-time systems to obtain the second order approximation suboptimal filter equation for nonlinear continuous-time systems.

## II. FILTERS FOR LINEAR DISCRETE-TIME SYSTEMS

### 1. The Dynamical System

The estimated linear discrete-time system can be described as:

$$X_{k+1} = \phi_k X_k + \Gamma_k W_k \quad (1-1)$$

$$(k = 0, 1, \dots)$$

$$Y_{k+1} = H_{k+1} X_{k+1} + V_{k+1} \quad (2-2)$$

Equation (2-1) is called a system equation.  $X \in \mathbb{R}^n$  is the state vector.  $\phi$  is an  $(n, n)$  transformation matrix.  $W \in \mathbb{R}^n$  is the noise vector.  $\Gamma$  is an  $(n, r)$  perturbation matrix. Equation (2-2) is called a monitoring equation.  $Y \in \mathbb{R}^m$  is the monitoring vector.  $H$  is an  $(m, n)$  monitoring matrix.  $V \in \mathbb{R}^m$  is the monitoring noise vector. If the system noise  $\{W_k\}$  and the monitoring noise  $\{V_k\}$  are unrelated Gaussian blank noise series with zero average value, with each  $k$  there is an  $l$  which:

$$EW_k = 0, EW_k W_l^T = Q_k \delta_{kl} \quad (Q_k \geq 0) \quad (2-3)$$

$$EV_k = 0, EV_k V_l^T = R_k \delta_{kl} \quad (R_k > 0) \quad (2-4)$$

$$EW_k V_l^T = 0 \quad (2-5)$$

Assuming that the initial state  $X_0$  is an  $n$ -dimensional vector, it is known that

$$EX_0 = \bar{X}_0, \quad E(X_0 - \bar{X}_0)(X_0 - \bar{X}_0)' = P_0 \quad (2-6)$$

If  $X_0$  is not related to  $\{W_k\}$  and  $\{V_k\}$ , then with respect to any  $k$ :

$$EX_k W_k' = 0, \quad EX_k V_k' = 0 \quad (2-7)$$

The block diagram of the estimated dynamical system is as shown in Figure 1 where  $D$  indicates a delay step.

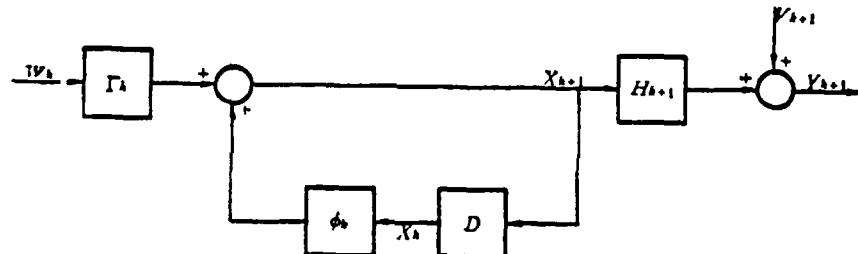


Figure 1. The Dynamic System (linear, discrete-time)

## 2. Filter Model

The state filter is capable of receiving monitoring signal  $V_{k+1}$  in real time and producing the estimated value  $\hat{X}_{k+1}$  for the state  $\hat{X}_{k+1}$  with the least or optimal estimated value error  $\tilde{X}_{k+1} = X_{k+1} - \hat{X}_{k+1}$  under the chosen guidelines.

Since the estimated state obeys the linear difference equation (2-1), it is only natural to consider that the state estimation also obeys a corresponding difference equation.

This indicates that the filter is also a linear discrete-time system which is capable of receiving monitoring signal  $Y_{k+1}$  and delivering the optimal value  $X_{k+1}$  for the state  $\hat{X}_{k+1}$ . Therefore, the structure of the filter can be assumed as shown in Figure 2. Its equations are:

$$Z_{k+1} = F_k Z_k + K_{k+1} Y_{k+1} \quad (2-8)$$

$(k=0, 1, \dots)$

$$\hat{X}_{k+1} = G_{k+1} Z_{k+1} \quad (2-9)$$

where  $Z \in R^n$ , is the filter state vector.  $\hat{X} \in R^m$ , is the filter output vector which is also the estimated value of state  $X$ .  $F$  is a  $(p,p)$  matrix which is the filter transformation matrix or the feedback matrix.  $K$  is a  $(p,m)$  matrix which is the filter front improvement matrix.  $G$  is a  $(n,p)$  matrix which is the filter output matrix.

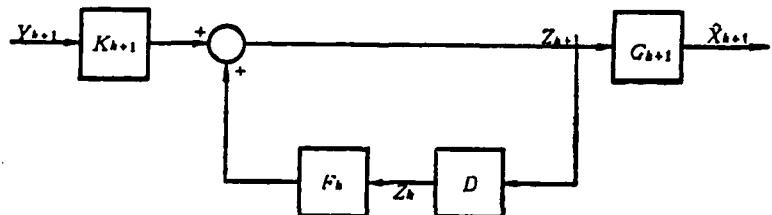


Figure 2. Filter Model (linear, discrete)

Once the structure of the filter is defined, the solution to the problem then relies on the determination of filter parameters and they are the dimension  $p$ , the feedback matrix  $F_k$ , front improvement matrix  $K_{k+1}$ , and output matrix  $G_{k+1}$ . To do that the guidelines must be pre-set. We have adopted the unbiased and minimum-variance rule which means when  $k = 0, 1, \dots$ .

$$(1) \text{ unbiased} \quad E\hat{X}_k = EX_k, \text{ 或 } E\hat{X}_k = 0 \quad (2-10)$$

$$(2) \text{ minimum variance} \quad \text{tr}P_k \triangleq \text{tr}E\hat{X}_k\hat{X}_k' \\ = \text{tr}E(X_k - \hat{X}_k)(X_k - \hat{X}_k)' = \min \quad (2-11)$$

### 3. Unbiased Filter

From the requirements of an unbiased filter, we shall determine parameters of the filter and the relationship between these parameters themselves. The first to be determined is the dimension of the filter  $p$ . At the initial moment  $k = 0$ , due to the requirements of an unbiased filter,  $E\hat{X}_0 = EX_0 = \hat{X}_0$ . However, the initial estimated value  $\hat{X}_0$  is not a stochastic quantity, therefore

$$\hat{X}_0 = \bar{X}_0 \quad (2-12)$$

On the other hand, the initial state of the filter  $Z_0$  should satisfy  $G_0 Z_0 = \hat{X}_0 = \bar{X}_0$ . In order to make the solution to  $Z_0$  unique, then  $p = n$  and  $\text{rank } G_0 = n$ . To further simplify the filter, let's make  $G_k = I_n$  ( $k = 0, 1, \dots$ ). Henceforth, the filter equations (2-8) and (2-9) become

$$\hat{X}_{k+1} = F_k \hat{X}_k + K_{k+1} Y_{k+1} \quad (2-13)$$

The initial value is  $\hat{X}_0 = X_0$ .

From the unbiased nature, the relationship between  $K_{k+1}$  and  $F_k$  can be determined. It can be derived from equations (2-1), (2-2), and (2-13) that

$$\begin{aligned} \hat{X}_{k+1} &= \phi_k X_k + \Gamma_k W_k - F_k \hat{X}_k - K_{k+1} (H_{k+1} \phi_k X_k + H_{k+1} \Gamma_k W_k + V_{k+1}) \\ &= (\phi_k - F_k - K_{k+1} H_{k+1} \phi_k) X_k + F_k \hat{X}_k + (I - K_{k+1} H_{k+1}) \Gamma_k W_k \\ &\quad - K_{k+1} V_{k+1} \end{aligned} \quad (2-14)$$

The unbiased filter's requirement is  $E\hat{X}_{k+1} = 0$ . The average values of  $\{W_k\}$  and  $\{V_k\}$  are zero. We found that

$$(\phi_k - F_k - K_{k+1} H_{k+1} \phi_k) E X_k = 0 \quad (2-15)$$

On the other hand, from (2-1) and (2-3) one can get

$$E X_{k+1} = \phi_k E X_k$$

The initial value is  $E X_0 = \bar{X}_0$ . Usually  $E X_k \neq 0$ ; therefore, equation (2-15) is valid for all  $k = 0, 1, \dots$ :

$$\phi_k - F_k - K_{k+1} H_{k+1} \phi_k = 0$$

The relationship between  $K_{k+1}$  and  $F_k$  is thus obtained:

$$F_k = \phi_k - K_{k+1} H_{k+1} \phi_k \quad (2-16)$$

Substituting (2-16) into (2-13), the filter equation becomes

$$\hat{X}_{k+1} = \phi_k \hat{X}_k + K_{k+1} (Y_{k+1} - H_{k+1} \phi_k \hat{X}_k) \quad (2-17)$$

with initial value  $\hat{X}_0 = \bar{X}_0$ . This filter is unbiased. Its block diagram is shown in Figure 3.

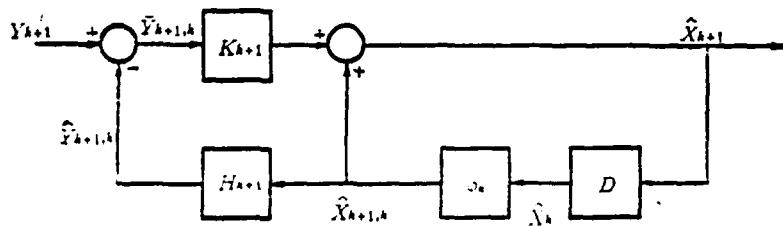


Figure 3. Unbiased Filter (linear, discrete)

From figures 1 and 3 we notice the characteristics of the structure of filters. First of all, a filter is a model of

the estimated system. Next, we note that filters come with a feedback mechanism which reflects the relationship of signal transformation. Before the newly monitored value  $Y_{k+1}$  is acquired, the filter provides an estimated value  $X_{k+1} = \phi_k \hat{X}_k$  state  $Y_{k+1}$  which is called the prediction state. The filter also provides the estimated value  $P_{k+1,k} = H_{k+1} \hat{X}_{k+1,k} = H_{k+1} \phi_k \hat{X}_k$ , of the measurable quantity  $Y_{k+1}$  which is defined as the predicted measurable quantity value. The filter uses the difference between  $P_{k+1,k}$  and  $P_{k+1,k} = Y_{k+1} - P_{k+1,k}$ , and the matrix  $K_{k+1}$ , to carry out linear "calibrations". It is not difficult to derive that the prediction state and the predicted measurable quantity value are unbiased:

$$E\hat{X}_{k+1,k} = EX_{k+1}, \quad E\hat{X}_{k+1,k} = 0 \quad (2-18)$$

$$E\hat{Y}_{k+1,k} = EV_{k+1}, \quad E\hat{Y}_{k+1,k} = 0 \quad (2-19)$$

Therefore, the unbiased filter equation (2-17) can be re-written as:

$$\hat{X}_{k+1} = \hat{X}_{k+1,k} + K_{k+1}(Y_{k+1} - H_{k+1}\hat{X}_{k+1,k}) \quad (2-20)$$

$$\hat{X}_{k+1,k} = \phi_k \hat{X}_k \quad (2-21)$$

with initial value  $\hat{X}_0 = \bar{X}_0$ .

#### 4. Unbiased and Minimum-Variance Filters

The remaining problem is to determine the matrix  $K_{k+1}$ , based on the minimum-variance requirement. Let's assume that  $X_k$  is optimum unbiased, and a minimum-variance value of the state  $X_k$ , the optimization matrix  $K_{k+1}$  in (2-17) or (2-20) can be determined based on (2-11).

By first calculating  $P_{k+1,k} \triangleq E \tilde{X}_{k+1,k} \tilde{X}_{k+1,k}^T$ , it is possible to obtain the following from (2-1) and (2-21):

$$\tilde{X}_{k+1,k} = X_{k+1} - \tilde{X}_{k+1,k} = \phi_k \tilde{X}_k + \Gamma_k W_k \quad (2-22)$$

Taking (2-3) and (2-7) into consideration, we get

$$P_{k+1,k} = \beta_k P_k \phi_k^T + \Gamma_k Q_k \Gamma_k^T \quad (2-23)$$

From (2-14), we can calculate  $P_{k+1,k}$ . Considering the unbiased filter requirement (2-16) with (2-22), we get

$$\begin{aligned} \tilde{X}_{k+1} &= (I - K_{k+1} H_{k+1}) \phi_k \tilde{X}_k + (I - K_{k+1} H_{k+1}) \Gamma_k W_k - K_{k+1} V_{k+1} \\ &= (I - K_{k+1} H_{k+1}) \tilde{X}_{k+1,k} - K_{k+1} V_{k+1} \end{aligned} \quad (2-24)$$

From (2-4), (2-5), (2-7), the difference equation obeyed by the error matrix of the unbiased filter is obtained as:

$$P_{k+1} = (I - K_{k+1} H_{k+1}) P_{k+1,k} (I - K_{k+1} H_{k+1})^T + K_{k+1} R_{k+1} K_{k+1}^T \quad (2-25)$$

with initial value  $P_0$ .

From  $\frac{\partial \text{tr} P_{k+1}}{\partial K_{k+1}} = 0$ , and the symmetry between  $P_{k+1,k}$  and  $R_{k+1}$  it is possible to obtain the necessary (as well as sufficient) condition of the optimum matrix:

$$-2P_{k+1,k} H_{k+1}^T + 2K_{k+1} (H_{k+1} P_{k+1,k} H_{k+1}^T + R_{k+1}) = 0$$

or

$$K_{k+1} (H_{k+1} P_{k+1,k} H_{k+1}^T + R_{k+1}) = P_{k+1,k} H_{k+1}^T \quad (2-26)$$

Since  $R_{k+1} > 0$ ,  $K_{k+1}$  becomes:

$$K_{k+1} = P_{k+1,k} H_{k+1}^T (H_{k+1} P_{k+1,k} H_{k+1}^T + R_{k+1})^{-1} \quad (2-27)$$

From (2-25), we get

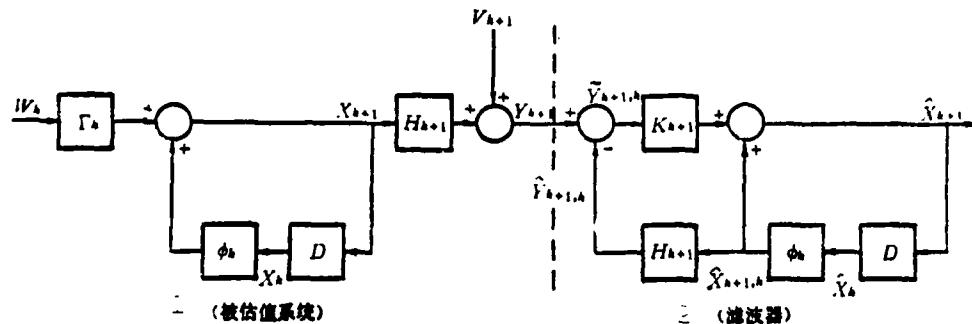
$$P_{k+1} = (I - K_{k+1} H_{k+1}) P_{k+1,k} - P_{k+1,k} H_{k+1} K_{k+1} + K_{k+1} (H_{k+1} P_{k+1,k} H_{k+1}^T + R_{k+1}) K_{k+1}^T$$

Substituting (2-26) into the above equation, we get the difference equation obeyed by the unbiased and minimum-variance filter:

$$P_{k+1} = (I - K_{k+1} H_{k+1}) P_{k+1,k} \quad (2-28)$$

Combining (2-20), (2-21), (2-23), (2-27), and (2-28), the recursive equation of the unbiased and minimum variance filter holds for a linear, discrete-time system. Therefore, we have derived the Kalman filter equation using a different approach. Based on the same concepts, we shall expand our treatment to the design of suboptimal filters for nonlinear, discrete-time systems.

Combining Figures 1 and 3, the block diagram of the dynamical system and the filter is as follows:



1. Dynamical System
2. Filter

Figure 4. Dynamical System and Filters (linear, discrete)

#### IV. FILTERS FOR NONLINEAR DISCRETE-TIME SYSTEMS

##### 1. The Evaluated System

The nonlinear discrete-time system can be described by the following equations:

$$X_{k+1} = \varphi(X_k, k) + \Gamma_k W_k \quad (3-1)$$

( $k = 0, 1, \dots$ )

$$Y_{k+1} = h(X_{k+1}, k+1) + V_{k+1} \quad (3-2)$$

where  $\varphi(\cdot)$  and  $h(\cdot)$  are  $n$  and  $m$  dimensional nonlinear vector functions, respectively. Other than these, the rest of the expressions are identical to the linear discrete-time system as described by (2-1) and (2-2).

##### 2. Filter Model

Based on the dynamical system (original) and the filter (model) and their corresponding relationship, the block diagram of the filter (including the dynamical system) can be obtained by referring to the linear discrete-time system as shown in Figure 4. It is illustrated in Figure 5.

##### 1. The dynamical system

##### 2. Filter

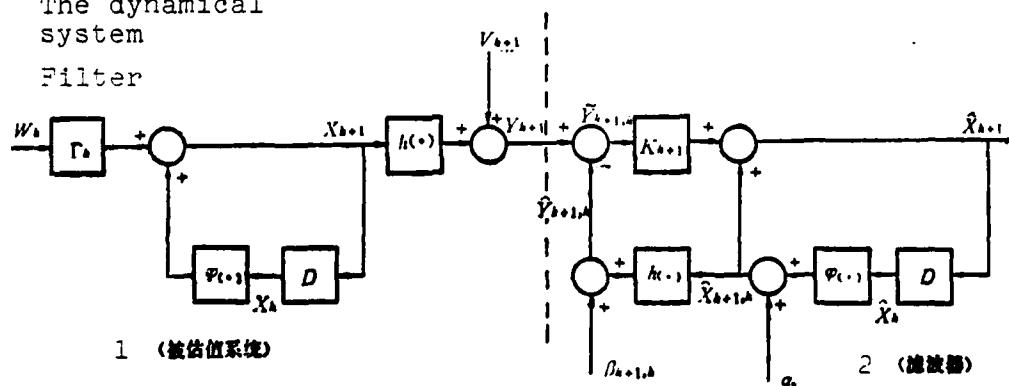


Figure 5. The System and the Filter (nonlinear, discrete-time)

Comparing Figure 5 with Figure 4, it is found that the filter for a nonlinear discrete-time system contains two external interaction terms  $\alpha_k$  and  $\beta_{k+1,k}$ . They represent the biased compensation terms of the estimated state and the monitored quantity, respectively.

The filter for a nonlinear discrete-time system obeys the following equations:

$$\hat{X}_{k+1} = \hat{X}_{k+1,k} + K_{k+1}[Y_{k+1} - h(\hat{X}_{k+1,k}, k+1) - \beta_{k+1,k}] \quad (3-3)$$

$$\hat{X}_{k+1,k} = \varphi(\hat{X}_k, k) + \alpha_k \quad (3-4)$$

with initial value  $\hat{X}_0 = \bar{X}_0$ .

### 3. Bias Compensation

From the requirement of unbiased filters, it is possible to approximately determine the bias compensation terms  $\alpha_k$  and  $\beta_{k+1,k}$ .

Assuming the former estimated value  $\hat{X}_k$  is unbiased, which means  $E\hat{X}_k = 0$ , the present unbiased estimated value of the state  $\hat{X}_{k+1,k}$  follows  $E\hat{X}_{k+1,k} = 0$ . From (3-1) and (3-4), we get

$$\hat{X}_{k+1,k} = [\varphi(X_k, k) - \varphi(\hat{X}_k, k)] + \Gamma_k W_k - \alpha_k \quad (3-5)$$

Since  $E\hat{X}_{k+1,k} = 0$ , the bias compensation term of the estimate state is:

$$\alpha_k = E[\varphi(X_k, k) - \varphi(\hat{X}_k, k)] \quad (3-6)$$

Unfortunately, the solution to (3-6) cannot be precisely determined. Therefore, only approximate values of  $a_k$  can be obtained. When Taylor expansion is carried out for the nonlinear function  $\varphi(\cdot)$  about  $\hat{X}_k$  the first order approximation of the bias compensation term of the estimated state is zero if only linear terms are taken into account. In fact:

$$a_k \approx E[\varphi(\hat{X}_k, k) + \frac{\partial \varphi_k}{\partial \hat{X}_k} (X_k - \hat{X}_k) - \varphi(\hat{X}_k, k)] = \frac{\partial \varphi_k}{\partial \hat{X}_k} E \tilde{X}_k = 0$$

where

$$\frac{\partial \varphi_k}{\partial \hat{X}_k} \triangleq \frac{\partial \varphi(X_k, k)}{\partial X_k} \Big|_{X_k = \hat{X}_k} \triangleq \begin{bmatrix} \frac{\partial \varphi_k^{(1)}}{\partial x_k^{(1)}} \dots \frac{\partial \varphi_k^{(1)}}{\partial x_k^{(n)}} \\ \vdots \quad \vdots \\ \frac{\partial \varphi_k^{(n)}}{\partial x_k^{(1)}} \dots \frac{\partial \varphi_k^{(n)}}{\partial x_k^{(n)}} \end{bmatrix} \Big|_{X_k = \hat{X}_k} \quad (3-7)$$

The  $(n, n)$  matrix,  $\frac{\partial \varphi}{\partial X}$ , is called the Jacobi matrix of the vector function  $\varphi(\cdot)$ .

To accurately determine the compensation, we expand the nonlinear function  $\varphi(\cdot)$  using Taylor expansion around  $X_k$  and then obtain the second order term. The second order approximation of the bias compensation term of the estimated state is then:

$$\begin{aligned} a_k &\approx E[\varphi(\hat{X}_k, k) + \frac{\partial \varphi_k}{\partial \hat{X}_k} \tilde{X}_k + \frac{1}{2} \sum_{i=1}^n (\text{tr} \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} \tilde{X}_k \tilde{X}_k^i) e_i - \varphi(\hat{X}_k, k)] \\ &= \frac{1}{2} \sum_{i=1}^n (\text{tr} \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} P_k) e_i \end{aligned} \quad (3-8)$$

where  $e_i$  ( $i = 1, 2, \dots, n$ ) is the natural base of the  $R^n$  space and

$$\frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} \triangleq \frac{\partial^2 \varphi^{(i)}(X_k, k)}{\partial X_k^2} \Big|_{X_k = \hat{X}_k} \triangleq \begin{bmatrix} \frac{\partial^2 \varphi_k^{(1)}}{\partial x_k^{(1)} \partial x_k^{(1)}} \dots \frac{\partial^2 \varphi_k^{(1)}}{\partial x_k^{(1)} \partial x_k^{(n)}} \\ \vdots \quad \vdots \\ \frac{\partial^2 \varphi_k^{(n)}}{\partial x_k^{(n)} \partial x_k^{(1)}} \dots \frac{\partial^2 \varphi_k^{(n)}}{\partial x_k^{(n)} \partial x_k^{(n)}} \end{bmatrix} \Big|_{X_k = \hat{X}_k} \quad (3-9)$$

The  $(n, n)$  matrix  $\frac{\partial^2 \varphi^{(i)}}{\partial X^2}$  is called the Hesse matrix (a symmetrical matrix) of the  $i$ th component  $\varphi^{(i)}(\cdot)$  of the vector function  $(\cdot)$ .

Furthermore, our required estimation of the measured quantity  $\hat{P}_{k+1,k}$  is also unbiased (i.e.,  $E\hat{P}_{k+1,k} = 0$ ). From (3-2) and (3-3) we get

$$\hat{Y}_{k+1,k} = [h(X_{k+1}, k+1) - h(\hat{X}_{k+1,k}, k+1)] + V_{k+1} - \beta_{k+1,k} \quad (3-10)$$

Now that  $E\hat{Y}_{k+1,k} = 0$ , the bias compensation term becomes:

$$\beta_{k+1,k} = E[h(X_{k+1}, k+1) - h(\hat{X}_{k+1,k}, k+1)] \quad (3-11)$$

Similarly, (3-11) cannot be solved precisely. Only approximations of  $\beta_{k+1,k}$  can be obtained. If a nonlinear function  $h(\cdot)$  undergoes Taylor expansion around  $\hat{X}_{k+1,k}$  the first order approximation of the bias compensation term of the predicated measured value is 0. In practice, because the estimated state  $\hat{X}_{k+1,k}$  after compensation  $\alpha_k$  is already almost unbiased, therefore

$$\begin{aligned} \beta_{k+1,k} &\approx E[h(\hat{X}_{k+1,k}, k+1) + \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}}(X_{k+1} - \hat{X}_{k+1,k}) - h(\hat{X}_{k+1,k}, k+1)] \\ &= \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} E\hat{X}_{k+1,k} \approx 0 \end{aligned}$$

where

$$\frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \triangleq \frac{\partial h(X_{k+1}, k+1)}{\partial X_{k+1}} \Big|_{X_{k+1} = \hat{X}_{k+1,k}} \triangleq \begin{bmatrix} \frac{\partial h_{k+1}^{(1)}}{\partial x_{k+1}^{(1)}} \dots \frac{\partial h_{k+1}^{(1)}}{\partial x_{k+1}^{(n)}} \\ \vdots \\ \frac{\partial h_{k+1}^{(m)}}{\partial x_{k+1}^{(1)}} \dots \frac{\partial h_{k+1}^{(m)}}{\partial x_{k+1}^{(n)}} \end{bmatrix} \Big|_{X_{k+1} = \hat{X}_{k+1,k}} \quad (3-12)$$

The  $(m, n)$  matrix  $\frac{\partial h}{\partial x}$  is called the Jacobi matrix of the vector function  $h(\cdot)$ .

Based on these results, both compensation terms  $a_k$  and  $b_{k+1,k}$  of the predicted state value and the estimated value of the measured quantity are 0 for first order nonlinear filters. This is an extrapolation of the Kalman filter.

In order to more precisely determine the compensation terms of the estimated value of the measured quantity, the nonlinear function  $h(\cdot)$  is similarly expanded about  $\hat{X}_{k+1,k}$ , using the Taylor expansion method to obtain the second order term. The second order approximation of the bias compensation term of the predicted value then becomes:

$$\begin{aligned}\beta_{k+1,k} &\approx E[h(\hat{X}_{k+1,k}, k+1) + \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \hat{X}_{k+1,k} \\ &\quad + \frac{1}{2} \sum_{j=1}^m (\text{tr} \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} \hat{X}_{k+1,k} \hat{X}_{k+1,k}^T) e_j - h(\hat{X}_{k+1,k}, k+1)] \\ &= \frac{1}{2} \sum_{j=1}^m (\text{tr} \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} P_{k+1,k}) e_j\end{aligned}\quad (3-13)$$

where  $e_j$  ( $j = 1, 2, \dots, m$ ) are the natural bases of the  $\mathbb{R}^n$  space and

$$\frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} \Delta \frac{\partial^2 h^{(j)}(X_{k+1}, k+1)}{\partial X_{k+1}^2} \Big|_{X_{k+1} = \hat{X}_{k+1,k}} \Delta \begin{bmatrix} \frac{\partial^2 h_{k+1}^{(j)}}{\partial x_{k+1}^{(1)} \partial x_{k+1}^{(1)}} & \dots & \frac{\partial^2 h_{k+1}^{(j)}}{\partial x_{k+1}^{(1)} \partial x_{k+1}^{(n)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 h_{k+1}^{(j)}}{\partial x_{k+1}^{(n)} \partial x_{k+1}^{(1)}} & \dots & \frac{\partial^2 h_{k+1}^{(j)}}{\partial x_{k+1}^{(n)} \partial x_{k+1}^{(n)}} \end{bmatrix} X_{k+1} = \hat{X}_{k+1,k} \quad (3-14)$$

The  $(n,n)$  matrix  $\frac{\partial^2 h^{(j)}}{\partial X^2}$  is called the Hesse matrix (symmetric matrix) of the  $j$ th component  $h^{(j)}(\cdot)$  of the vector function  $h(\cdot)$ .

After making the approximate compensations  $a_k$  and  $\beta_{k+1,k}$  the next state estimation  $X_{k+1}$  is almost unbiased. From (3-3) we get:

$$E \hat{X}_{k+1} = E(\hat{X}_{k+1,k} - K_{k+1} \hat{Y}_{k+1,k}) = E \hat{X}_{k+1,k} - K_{k+1} E \hat{Y}_{k+1,k} \approx 0$$

Thus, we obtain the second order approximation of the unbiased filters for nonlinear discrete-time systems.

#### 4. Suboptimal Unbiased and Minimum Variance Filters

Based on the requirement of minimum variance, we are going to determine the second order approximation of the suboptimal matrix  $K_{k+1}$  of unbiased filters. For this purpose, let's assume that  $\{\bar{X}_k\}$  and  $\{\bar{X}_{k+1,k}\}$  are Gaussian. For a zero average Gaussian series, it is known from references [5,6] that the odd order terms are zero. For example, if  $\{\xi_i\}$  is a one-dimensional zero average Gaussian stochastic series, then for any  $i, j, k =$

$$E\xi_i \xi_j \xi_k = 0 \quad (3-15)$$

and for any  $i, j, k$ , and  $l$ :

$$E\xi_i \xi_j \xi_k \xi_l = (E\xi_i \xi_j)(E\xi_k \xi_l) + (E\xi_i \xi_k)(E\xi_j \xi_l) + (E\xi_i \xi_l)(E\xi_j \xi_k) \quad (3-16)$$

Let's calculate  $P_{k+1,k}$  first. Based on (3-4) and (3-8) we get:

$$\bar{X}_{k+1,k} = -\frac{\partial \varphi_k}{\partial \bar{X}_k} \bar{X}_k + \Gamma_k W_k + \frac{1}{2} \sum_{i=1}^n [\text{tr} \frac{\partial^2 \varphi_k^{(i)}}{\partial \bar{X}_k^2} (\bar{X}_k \bar{X}_k^T - P_k)] e_i \quad (3-17)$$

Comparing (3-17) with (2-22) and then taking (3-15) and (3-16) into account, we get

$$P_{k+1,k} = \frac{\partial \varphi_k}{\partial \bar{X}_k} P_k \left( \frac{\partial \varphi_k}{\partial \bar{X}_k} \right)' + \Gamma_k Q_k \Gamma_k' + J P_k \quad (3-18)$$

where

$$\begin{aligned} J P_k &= \frac{1}{4} E \left\{ \sum_{i=1}^n [\text{tr} \frac{\partial^2 \varphi_k^{(i)}}{\partial \bar{X}_k^2} (\bar{X}_k \bar{X}_k^T - P_k)] e_i \right\} \left\{ \sum_{j=1}^n [\text{tr} \frac{\partial^2 \varphi_k^{(j)}}{\partial \bar{X}_k^2} (\bar{X}_k \bar{X}_k^T - P_k)] e_j \right\}' \\ &= \frac{1}{2} E \left\{ \sum_{i,j=1}^n [\text{tr} \frac{\partial^2 \varphi_k^{(i)}}{\partial \bar{X}_k^2} (\bar{X}_k \bar{X}_k^T - P_k)] [\text{tr} \frac{\partial^2 \varphi_k^{(j)}}{\partial \bar{X}_k^2} (\bar{X}_k \bar{X}_k^T - P_k)] e_i e_j \right\} \end{aligned}$$

The  $(i,j)^{th}$  element of the  $(n,n)$  matrix  $\Delta P_k$  is:

$$\begin{aligned} (\Delta P_k)_{ij} &= \frac{1}{4} E \left[ \text{tr} \left( \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} (\hat{X}_k \hat{X}_k' - P_k) \right) \left[ \text{tr} \left( \frac{\partial^2 \varphi_k^{(j)}}{\partial \hat{X}_k^2} (\hat{X}_k \hat{X}_k' - P_k) \right) \right] \right] \\ &= \frac{1}{4} \left\{ E \left[ \text{tr} \left( \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} \hat{X}_k \hat{X}_k' \right) \text{tr} \left( \frac{\partial^2 \varphi_k^{(j)}}{\partial \hat{X}_k^2} \hat{X}_k \hat{X}_k' \right) \right] - \text{tr} \left( \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} P_k \right) \text{tr} \left( \frac{\partial^2 \varphi_k^{(j)}}{\partial \hat{X}_k^2} P_k \right) \right\} \end{aligned}$$

From (3-16) and the symmetry of the Hesse matrix, one gets:

$$(\Delta P_k)_{ij} = \frac{1}{2} \text{tr} \left( \frac{\partial^2 \varphi_k^{(i)}}{\partial \hat{X}_k^2} P_k \right) \left( \frac{\partial^2 \varphi_k^{(j)}}{\partial \hat{X}_k^2} P_k \right) \quad (3-19)$$

Let's calculate  $P_{k+1}$  now. From (3-3) we get

$$\hat{X}_{k+1} = X_{k+1} - \hat{X}_{k+1} = \hat{X}_{k+1,k} - K_{k+1} \bar{Y}_{k+1,k}$$

Plugging (3-10) and (3-12) into the above equation, we get:

$$\begin{aligned} \hat{X}_{k+1} &= \left( I - K_{k+1} \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right) \hat{X}_{k+1,k} \\ &\quad - K_{k+1} \left\{ V_{k+1} + \frac{1}{2} \sum_{j=1}^n \left[ \text{tr} \left( \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} (\hat{X}_{k+1,k} \hat{X}_{k+1,k}' - P_{k+1,k}) \right) e_j \right] \right\} \end{aligned} \quad (3-20)$$

Comparing (3-20) with (2-24) and noticing (3-15) and (3-16), we obtain the difference equation obeyed by the variance matrix of the second order approximation of the unbiased filter:

$$\begin{aligned} P_{k+1} &= \left( I - K_{k+1} \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right) P_{k+1,k} \left( I - K_{k+1} \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right)' + \\ &\quad + K_{k+1} (R_{k+1} + \Delta R_{k+1,k}) K_{k+1}' \end{aligned} \quad (3-21)$$

with initial value  $P_0$  where

$$\begin{aligned}
\mathcal{A}R_{k+1,k} &= \frac{1}{4} E \left\{ \sum_{i=1}^m \left[ \text{tr} \frac{\partial^2 h_{k+1}^{(i)}}{\partial \hat{X}_{k+1,k}^2} (\hat{X}_{k+1,k} \hat{X}_{k+1,k}' \right. \right. \\
&\quad \left. \left. - P_{k+1,k}) \right] e_i \right\} \left\{ \sum_{j=1}^m \left[ \text{tr} \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} (\hat{X}_{k+1,k} \hat{X}_{k+1,k}' - P_{k+1,k}) \right] e_j \right\}' \\
&= \frac{1}{4} E \left\{ \sum_{i,j=1}^m \left[ \text{tr} \frac{\partial^2 h_{k+1}^{(i)}}{\partial \hat{X}_{k+1,k}^2} (\hat{X}_{k+1,k} \hat{X}_{k+1,k}' \right. \right. \\
&\quad \left. \left. - P_{k+1,k}) \right] \left[ \text{tr} \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} (\hat{X}_{k+1,k} \hat{X}_{k+1,k}' - P_{k+1,k}) \right] e_i e_j' \right\}
\end{aligned}$$

Similar to the calculation of the  $(n,n)$  matrix  $\mathcal{A}P_k$ , it is possible to obtain the  $(i,j)^{\text{th}}$  element of the  $(m,m)$  matrix  $\mathcal{A}R_{k+1,k}$  which is

$$(\mathcal{A}R_{k+1,k})_{ij} = \frac{1}{2} \text{tr} \left( \frac{\partial^2 h_{k+1}^{(i)}}{\partial \hat{X}_{k+1,k}^2} P_{k+1,k} \right) \left( \frac{\partial^2 h_{k+1}^{(j)}}{\partial \hat{X}_{k+1,k}^2} P_{k+1,k} \right) \quad (3-22)$$

Comparing (3-21) with (3-25), we notice that they are identical in form. Therefore,  $K_{k+1}$  can be obtained using similar procedures:

$$K_{k+1} = P_{k+1,k} \left( \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right)' \left[ \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} P_{k+1,k} \left( \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right)' + R_{k+1} + \mathcal{A}R_{k+1,k} \right]^{-1} \quad (3-23)$$

Corresponding to (2-28), the variance matrix of the second order approximation of the suboptimal filters should satisfy the difference equation:

$$P_{k+1} = \left( I - K_{k+1} \frac{\partial h_{k+1}}{\partial \hat{X}_{k+1,k}} \right) P_{k+1,k} \quad (3-24)$$

Combining (3-3), (3-4), (3-8), (3-13), (3-18), (3-19), (3-22), (3-23), and (3-24), the recursive equation of the second order approximation, minimum-variance filter. Suboptimal filter of nonlinear discrete-time system can be obtained. In the above set of equations, if we let  $\mathcal{A}P_k$ ,  $\mathcal{A}R_{k+1,k}$  be small, it transforms into the first order approximation suboptimal filter. This is the well known expanded Kalman filter.

In engineering applications, if  $\mathcal{A}P_k$  is (3-18) is smaller than  $\mathcal{A}R_{k+1,k}$  in (3-23) is also smaller than  $\mathcal{A}P_k$ , they can be neglected to simplify calculation steps.

## IV. FILTERS OF LINEAR CONTINUOUS-TIME SYSTEMS

### 1. The Dynamical System

A linear continuous-time system can be described by:

$$\dot{X}(t) = A(t)X(t) + \Gamma(t)W(t) \quad (4-1)$$

$$Y(t) = C(t)X(t) + V(t) \quad (4-2)$$

Equation (4-1) is called the system equation.  $X \in \mathbb{R}^n$  is the state vector.  $A$  is a  $(n,n)$  matrix known as the system matrix.  $W \in \mathbb{R}^n$  is the system noise vector.  $\Gamma$  is a  $(n,r)$  matrix called the perturbation matrix. Equation (4-2) is called the monitoring equation.  $Y \in \mathbb{R}^m$  is the measuring vector.  $C$  is a  $(m,n)$  matrix called the measuring matrix.  $V \in \mathbb{R}^m$  is the measurement noise vector. Assuming that the system noise  $\{W(t)\}$  and the measurement noise  $\{V(t)\}$  are unrelated to zero average Gaussian blank noise processes, with respect to any  $t$ , there is

$$EW(t) = 0, \quad EW(t)W'(t) = Q(t)\delta(t-t) \quad (Q(t) \geq 0) \quad (4-3)$$

$$EV(t) = 0, \quad EV(t)V'(t) = R(t)\delta(t-t) \quad (R(t) \geq 0) \quad (4-4)$$

$$EW(t)V'(t) = 0 \quad (4-5)$$

Assuming further that the initial state  $X(t_0) = X_0$  is a  $n$  dimensional Gaussian stochastic vector, it has been shown that

$$EX_0 = \bar{X}_0, \quad E(X_0 - \bar{X}_0)(X_0 - \bar{X}_0)' = P_0 \quad (4-6)$$

$X_0$  and  $\{W(t)\}$  and  $\{V(t)\}$  are not related, i.e. with respect to any  $t$  there is

$$EX_0W'(t) = 0, EX_0V'(t) = 0$$

(4-7)

The block diagram of the dynamical system is shown in Figure 6.

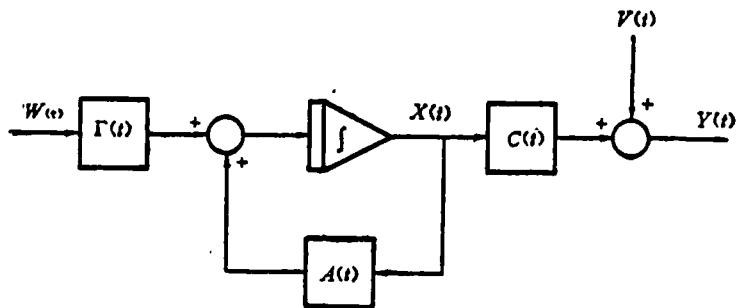


Figure 6. The Dynamical System (linear, continuous)

## 2. Filter Model

Since the state obeys the differential equation (4-1), it can be derived that the estimated value of the state should obey a corresponding linear differential equation. Therefore, the filter should also be a linear continuous time system. It can receive measurement signal Y(t) and provide the optimum estimated value  $\hat{X}(t)$  for state  $\hat{X}(t)$ . Henceforth, the structure of the filter is assumed as shown in Figure 7.

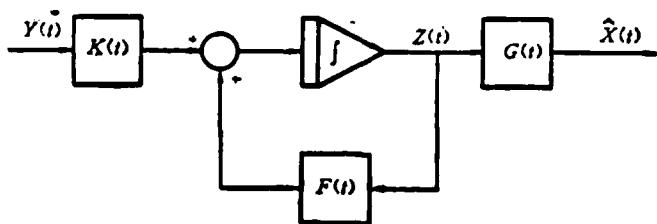


Figure 7. Filter Model (linear, continuous)

Its equations are:

$$\dot{Z}(t) = F(t)Z(t) + K(t)Y(t) \quad (4-8)$$

$$\dot{\hat{X}}(t) = G(t)Z(t) \quad (4-9)$$

where  $Z \in R^n$ , is the state vector of the filter. The filter output vector is the estimated value of state  $\hat{X} \in R^m$ ,  $F$  is a  $(n, p)$  matrix known as the filter system matrix or the feedback matrix.  $K$  is a  $(p, m)$  matrix called the filter front improvement matrix.  $G$  is a  $(n, p)$  matrix known as the filter output matrix.

Similar to the treatment of linear discrete-time systems, we are going to determine the dimension  $p$  of the filter, the feedback matrix  $F(t)$ , the matrix  $K(t)$ , and the output matrix  $G(t)$  based on the unbiased, minimum-variance guidelines. For all  $t \geq t_0$ , we get

1) unbiased relationship  $E\hat{X}(t) = EX(t)$ , or  $E\hat{X}(t) = 0$  (4-10)

2) minimum variance  $\text{tr} P(t) \triangleq \text{tr } E\hat{X}(t)\hat{X}'(t)$  (4-11)  
 $= \text{tr } E[X(t) - \hat{X}(t)][X(t) - \hat{X}(t)]' = \min$

### 3. Unbiased Filters

We'll first discuss the properties of unbiased filters here.

Same as the treatment used for linear discrete-time system, it is possible to obtain that the dimension of the filter  $p = n$  and  $G(t) = I_n (t \geq t_0)$ . Therefore, the filter equations (4-8) and (4-9) become:

$$\dot{\hat{X}}(t) = F(t)\hat{X}(t) + K(t)Y(t) \quad (4-12)$$

with initial value  $\bar{X}(t_0) = \bar{X}_0$ .

From the requirement of unbiased filters, the relationship between  $K(t)$  and the feedback matrix  $F(t)$  can be determined. From (4-1), (4-2) and (4-12) we get

$$\begin{aligned}\dot{\bar{X}}(t) &= A(t)X(t) + \Gamma(t)W(t) - F(t)\bar{X}(t) - K(t)[C(t)X(t) + V(t)] \\ &= [A(t) - F(t) - K(t)C(t)]X(t) + F(t)\bar{X}(t) + \Gamma(t)W(t) - K(t)V(t)\end{aligned}\quad (4-13)$$

Due to the unbiased property:  $E\bar{X}(t) = 0$ . Then  $\frac{d}{dt}E\bar{X}(t) = E\dot{\bar{X}}(t) = 0$  which leads to the fact that  $\{W(t)\}$  and  $\{V(t)\}$  are zero.

Therefore:

$$[A(t) - F(t) - K(t)C(t)]EX(t) = 0 \quad (4-14)$$

On the other hand, from (4-1) and (4-3) we get

$$E\dot{X}(t) = A(t)EX(t)$$

with initial value  $EX(t_0) = \bar{X}_0$ . Therefore, usually  $EX(t) \neq 0$ . But (4-14) is valid for all  $t \geq t_0$ .

$$A(t) - F(t) - K(t)C(t) = 0$$

The relationship between  $K(t)$  and  $F(t)$  for the filter is thus determined as:

$$F(t) = A(t) - K(t)C(t) \quad (4-15)$$

Substituting (4-15) into (4-12), the unbiased filter equation can be obtained:

$$\dot{\hat{X}}(t) = A(t)\hat{X}(t) + K(t)[Y(t) - C(t)\hat{X}(t)] \quad (4-16)$$

with initial value  $\hat{X}(t_0) = \hat{X}_0$ . Its block diagram is shown in Figure 8.

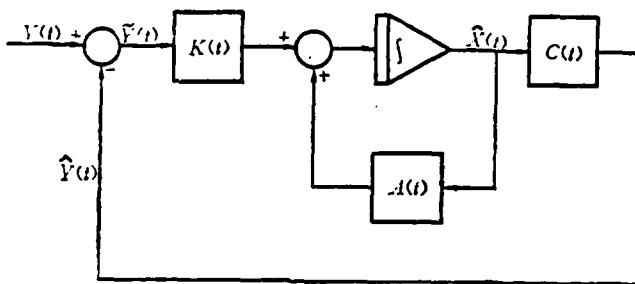


Figure 8. Unbiased Filter (linear, continuous)

Comparing Figure 8 with Figure 6, it can be found that the filter is a model of the dynamical system. Furthermore, the filter has a feedback structure which is capable of obtaining the difference  $\hat{P}(t) = C(t)\hat{X}(t)$  between the measured value  $Y(t)$  and the estimated value  $\hat{Y}(t) = Y(t) - \hat{P}(t)$  and it is a measure to determine the prediction error  $\hat{X}(t) = X(t) - \hat{X}(t)$ . It is then possible to use  $K(t)$  to carry out linear "calibrations".

#### 4. Unbiased Minimum-variance Filters

Let us now use the minimum-variance requirement to determine the matrix  $K(t)$  for unbiased filters. From (4-13) and the unbiased filter requirement (4-15), we get

$$\dot{\hat{X}}(t) = [A(t) - K(t)C(t)]\hat{X}(t) + \Gamma(t)W(t) - K(t)V(t) \quad (4-17)$$

In (4-17), let's assume the matrix  $\psi(t, t_0)$  is the transformation matrix with respect to the system matrix  $A(t) - K(t) C(t)$ ; then the solution becomes

$$\tilde{\mathbf{X}}(t) = \psi(t, t_0) \tilde{\mathbf{X}}(t_0) + \int_{t_0}^t \psi(t, \tau) [\Gamma(\tau) W(\tau) - K(\tau) V(\tau)] d\tau \quad (4-18)$$

Noticing (4-3), (4-4), (4-5), (4-6), and (4-7), we get

$$P(t) \triangleq E \tilde{\mathbf{X}}(t) \tilde{\mathbf{X}}'(t) = \psi(t, t_0) P_0 \psi'(t, t_0) + \int_{t_0}^t \psi(t, \tau) [\Gamma(\tau) Q(\tau) \Gamma'(\tau) + K(\tau) R(\tau) K'(\tau)] \psi'(\tau, \tau) d\tau$$

It is the solution to the following matrix equation

$$\begin{aligned} \dot{P}(t) = & [A(t) - K(t) C(t)] P(t) + P(t) [A(t) - K(t) C(t)]' + \\ & + \Gamma(t) Q(t) \Gamma'(t) + K(t) R(t) K'(t) \end{aligned} \quad (4-19)$$

with initial value  $P(t_0) = P_0$ .

In order to determine  $K(t)$ , we consider (4-19) as the differential equation obeyed by the dynamical system. The state variable of the dynamical system is the element  $p_{ij}(t)$ , of the  $(n, n)$  matrix  $P(\cdot)$ . The control variable is the element  $k_{ij}(t)$ , of the  $(n, m)$  matrix  $K(t)$ . The minimum-variance guideline is

$$J = \text{tr } P(T) = \min, \quad (t_0 \leq t \leq T) \quad (4-20)$$

In summary, the problem becomes the determination of the matrix  $K(t)$  for a given system equation (4-19) with terminal time  $T$  and characteristic function (4-20) under which  $J = \min$ . This becomes a classical optimal control problem for a fixed type. We will solve it using the following method. Let's assume the optimal function is:

$$V[P(t), t] = \min_{\{K(\tau), t \leq \tau \leq T\}} J = \min_{\{K(\tau), t \leq \tau \leq T\}} \text{tr} P(T) \quad (4-21)$$

The Hamilton-Jacobi matrix equation is

$$\text{tr} \frac{\partial V}{\partial P} P = - \frac{\partial V}{\partial t} \quad (4-22)$$

Since  $\frac{\partial V}{\partial t}$  is not related to  $K(t)$ , we get

$$\frac{\partial}{\partial K(t)} \text{tr} \frac{\partial V}{\partial P} P = 0 \quad (4-23)$$

Plugging (4-19) into (4-23) and keeping in mind that  $\frac{\partial V}{\partial P}$  is unrelated to  $K(t)$ , we get

$$2 \frac{\partial V}{\partial P} [P(t)C'(t) - K(t)R(t)] = 0$$

Since  $\frac{\partial V}{\partial P}$  is orthogonal [3], the necessary (and sufficient) condition for the matrix  $K(t)$ :

$$K(t)R(t) = P(t)C'(t) \quad (4-24)$$

Due to the fact that  $R(t)$  is orthogonal,  $K(t)$  thus becomes:

$$K(t) = P(t)C'(t)R^{-1}(t) \quad (4-25)$$

From (4-19), we get

$$\begin{aligned} \dot{P}(t) &= A(t)P(t) + P(t)A'(t) + \Gamma(t)Q(t)\Gamma'(t) \\ &\quad - K(t)C(t)P(t) - P(t)C'(t)K'(t) + K(t)R(t)K'(t) \end{aligned}$$

Substituting (4-24) into the above equation, the differential equation which describes the error matrix of the unbiased minimum-variance filter is obtained:

$$\dot{P}(t) = A(t)P(t) + P(t)A'(t) + \Gamma(t)Q(t)\Gamma'(t) - K(t)R(t)K'(t) \quad (4-26)$$

with initial value  $P(t_0) = P_0$ .

Combining (4-16), (4-25), and (4-26), we obtain the complete equation of the unbiased minimum-variance filter of a linear continuous-time system. Therefore, we derived the Kalman-Bucy equation from a different approach. Along the same idea, we will extrapolate our treatment to nonlinear continuous-time systems to obtain the suboptimal filter.

Combining Figures 6 and 8, the block diagram of the filter and the dynamical system is as follows:

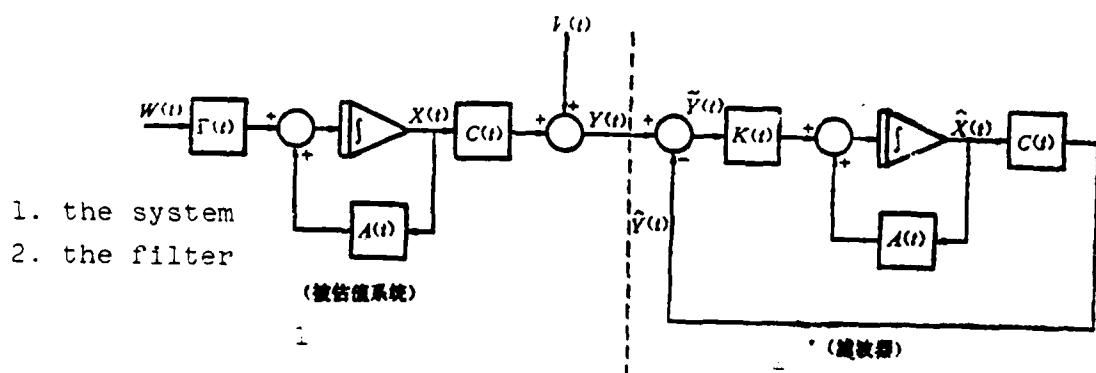


Figure 9. The System and the Filter (nonlinear, continuous)

## V. FILTERS FOR NONLINEAR CONTINUOUS-TIME SYSTEMS

### 1. The Estimated System

The estimated nonlinear continuous-time system can be described by:

$$\dot{\hat{X}}(t) = f[X(t), t] + \Gamma(t)W(t) \quad (5-1)$$

$$Y(t) = h[X(t), t] + V(t) \quad (5-2)$$

where  $f(\cdot)$  and  $h(\cdot)$  are  $n$  and  $m$  dimensional nonlinear vector functions, respectively. Other than that, the remaining are the same as (4-1) and (4-2) for linear continuous-time systems.

### 2. Filter Model

We use the corresponding relationship between the estimated system (original) and the filter (model). Referring to Figure 9 for linear continuous-time systems, we can determine the block diagram for the filter (including the system itself) as shown in Figure 10:

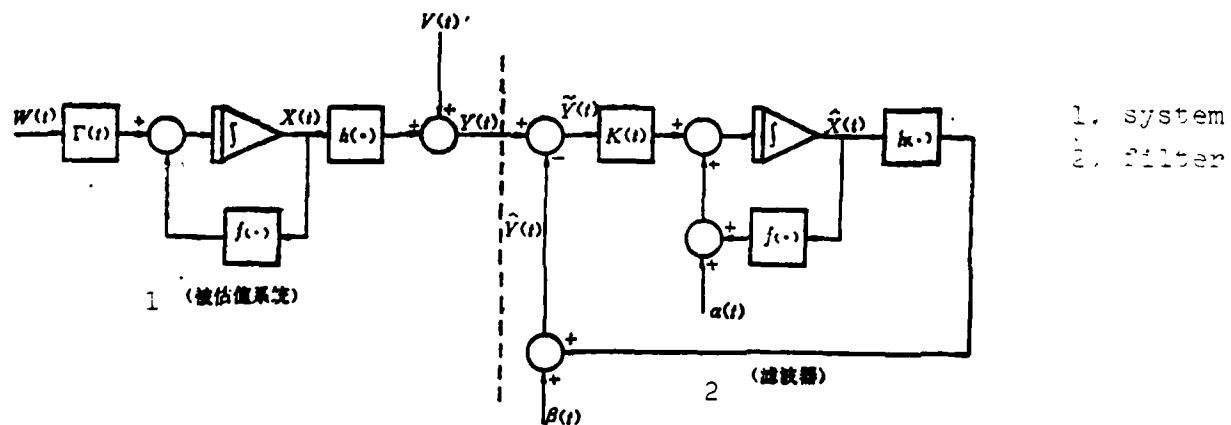


Figure 10. The System and the Filter (nonlinear, continuous)

Comparing Figure 9 with Figure 10, there are two external reaction terms  $\alpha(t)$  and  $\beta(t)$  added to the filters of nonlinear continuous-time systems. They are the bias compensation terms for the state and the measured quantity, respectively.

Henceforth, the equation of the filter of a nonlinear continuous-time system is:

$$\dot{\hat{X}}(t) = f[\hat{X}(t), t] + \alpha(t) + K(t)\{Y(t) - h[\hat{X}(t), t] - \beta(t)\} \quad (5-3)$$

with initial value  $\hat{X}(t_0) = \hat{X}_0$ .

### 3. Bias Compensation

From the unbias requirement, the bias compensation terms  $\alpha(t)$ ,  $\beta(t)$  can be determined. From (5-1), (5-2) and (5-3), we get

$$\begin{aligned} \dot{\hat{X}}(t) &= \{f[X(t), t] - f[\hat{X}(t), t] - \alpha(t)\} + F(t)W(t) \\ &\quad - K(t)\{h[X(t), t] - h[\hat{X}(t), t] - \beta(t)\} - K(t)V(t) \end{aligned} \quad (5-4)$$

The unbias requirement is  $E\hat{X}(t) = 0$ . Therefore,  $\frac{d}{dt}E\hat{X}(t) = E\dot{\hat{X}}(t) = 0$ . This makes the average values of  $\{W(t)\}$  and  $\{V(t)\}$  are 0. Based on these considerations, the state and measured quantity should be compensated by  $\alpha(t)$  and  $\beta(t)$ , respectively.

$$\alpha(t) = E\{f[X(t), t] - f[\hat{X}(t), t]\} \quad (5-5)$$

$$\beta(t) = E\{h[X(t), t] - h[\hat{X}(t), t]\} \quad (5-6)$$

However, the exact solutions to (5-5) and (5-6) cannot be determined. Therefore, only approximations of  $\alpha(t)$  and  $\beta(t)$  can be obtained.

If we carry out Taylor expansion around  $\hat{X}(t)$  for nonlinear functions  $f(\cdot)$  and  $h(\cdot)$  and take the linear terms, then the first order approximations of the bias compensation terms of the state and the measured quantity are zero.

In fact:

$$a(t) \cong E\{f[\hat{X}(t), t] + \frac{\partial f_t}{\partial \hat{X}(t)} \hat{X}(t) - f[\hat{X}(t), t]\} = \frac{\partial f_t}{\partial \hat{X}(t)} E \hat{X}(t) = 0$$

$$\beta(t) \cong E\{h[\hat{X}(t), t] + \frac{\partial h_t}{\partial \hat{X}(t)} \hat{X}(t) - h[\hat{X}(t), t]\} = \frac{\partial h_t}{\partial \hat{X}(t)} E \hat{X}(t) = 0$$

where the  $(n, n)$  matrix  $\frac{\partial f_t}{\partial \hat{X}(t)}$  and  $(m, n)$  matrix  $\frac{\partial h_t}{\partial \hat{X}(t)}$  are the values of the Jacobi matrix of the vector functions  $f(\cdot)$  and  $h(\cdot)$  at  $X(t) = \hat{X}(t)$ , respectively.

To obtain the compensation terms more precisely, we are going to expand the nonlinear functions  $f(\cdot)$  and  $h(\cdot)$  around  $\hat{X}(t)$  using Taylor expansion and include second order terms, the second order approximations of the bias compensation terms of the state and the measured quantity are:

$$\begin{aligned} a(t) &\cong E\left\{f[\hat{X}(t), t] + \frac{\partial f_t}{\partial \hat{X}(t)} \hat{X}(t) + \frac{1}{2} \sum_{i=1}^n \left[ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \hat{X}^i(t)} \hat{X}(t) \hat{X}'(t) \right] e_i - f[\hat{X}(t), t] \right\} \\ &= \frac{1}{2} \sum_{i=1}^n \left[ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \hat{X}^i(t)} P(t) \right] e_i \end{aligned} \quad (5-7)$$

$$\begin{aligned} \beta(t) &\cong E\left\{h[\hat{X}(t), t] + \frac{\partial h_t}{\partial \hat{X}(t)} \hat{X}(t) + \frac{1}{2} \sum_{i=1}^m \left[ \text{tr} \frac{\partial^2 h_t^{(i)}}{\partial \hat{X}^i(t)} \hat{X}(t) \hat{X}'(t) \right] e_i - h[\hat{X}(t), t] \right\} \\ &= \frac{1}{2} \sum_{i=1}^m \left[ \text{tr} \frac{\partial^2 h_t^{(i)}}{\partial \hat{X}^i(t)} P(t) \right] e_i \end{aligned} \quad (5-8)$$

where the  $(n, n)$  matrix  $\frac{\partial^2 f_t^{(i)}}{\partial \hat{X}^i}$  and  $\frac{\partial^2 h_t^{(i)}}{\partial \hat{X}^i}$  are the values of the Hesse matrix of the  $i^{\text{th}}$  component  $f^{(i)}(\cdot)$  of the vector function  $f(\cdot)$  and the  $j^{\text{th}}$  component  $h^{(j)}(\cdot)$  of  $h(\cdot)$  at  $X(t) = \hat{X}(t)$  and they are symmetric.

Henceforth, we have obtained the second order approximation unbiased filter for nonlinear continuous-time systems.

#### 4. Unbiased Minimum-Variance Suboptimal Filter

Now we are going to determine the second order approximation suboptimal matrix  $K(t)$  of the unbiased filter. Similar to the discrete-time systems, we assume that  $\{\bar{X}(t)\}$  is almost a Gaussian process. Substituting (5-7) and (5-8) into (5-4), we get

$$\begin{aligned}\dot{\bar{X}}'(t) \simeq & \left[ \frac{\partial f_t}{\partial \bar{X}(t)} - K(t) \frac{\partial h_t}{\partial \bar{X}(t)} \right] \bar{X}(t) + \Gamma(t) W(t) - K(t) V(t) + \\ & + \frac{1}{2} \sum_{i=1}^n \left\{ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \bar{X}^2(t)} [\bar{X}(t) \bar{X}'(t) - P(t)] \right\} e_i - \\ & - \frac{1}{2} K(t) \sum_{i=1}^n \left\{ \text{tr} \frac{\partial^2 h_t^{(i)}}{\partial \bar{X}^2(t)} [\bar{X}(t) \bar{X}'(t) - P(t)] \right\} e_i\end{aligned}\quad (5-9)$$

and

$$\dot{P}(t) = \frac{d}{dt} E \bar{X}(t) \bar{X}'(t) = E [\dot{\bar{X}}(t) \bar{X}'(t) + \bar{X}(t) \dot{\bar{X}}'(t)] \quad (5-10)$$

Substituting (5-9) into (5-10) and assuming  $\{\bar{X}(t)\}$  approximates a zero average Gaussian process, we get

$$\begin{aligned}E \left\{ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \bar{X}^2(t)} [\bar{X}(t) \bar{X}'(t) - P(t)] \right\} \bar{X}'(t) \\ = E \left[ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \bar{X}^2(t)} \bar{X}(t) \bar{X}'(t) \right] \bar{X}'(t) - \left[ \text{tr} \frac{\partial^2 f_t^{(i)}}{\partial \bar{X}^2(t)} P(t) \right] E \bar{X}'(t) = 0\end{aligned}$$

similarly

$$E \left\{ \text{tr} \frac{\partial^2 h_t^{(i)}}{\partial \bar{X}^2(t)} [\bar{X}(t) \bar{X}'(t) - P(t)] \right\} \bar{X}'(t) = 0$$

(5-10) can be reexpressed as:

$$\begin{aligned} \dot{P}(t) = & \left[ -\frac{\partial f_t}{\partial \hat{X}(t)} - K(t) \frac{\partial h_t}{\partial \hat{X}(t)} \right] P(t) + P(t) \left[ \frac{\partial f_t}{\partial \hat{X}(t)} - K(t) \frac{\partial h_t}{\partial \hat{X}(t)} \right]' + \\ & + 2E[\Gamma(t)W(t)\tilde{X}'(t)] - 2E[K(t)V(t)\tilde{X}'(t)] \end{aligned} \quad (5-11)$$

In the calculation of  $E[T(g)W(t)\tilde{X}'(t)]$  and  $E[k(t)V(t)\tilde{X}'(t)]$ , more simplification can be made by neglecting the second order terms in (5-9). Then  $\dot{\tilde{X}}(t)$  is simplified as

$$\dot{\tilde{X}}(t) \equiv \left[ \frac{\partial f_t}{\partial \hat{X}(t)} - K(t) \frac{\partial h_t}{\partial \hat{X}(t)} \right] \tilde{X}(t) + \Gamma(t)W(t) - K(t)V(t) \quad (5-12)$$

Henceforth, it is not too difficult from (5-11) to obtain the following using a method analogous to that of the linear continuous systems.

$$\begin{aligned} \dot{P}(t) = & \left[ \frac{\partial f_t}{\partial \hat{X}(t)} - K(t) \frac{\partial h_t}{\partial \hat{X}(t)} \right] P(t) + P(t) \left[ \frac{\partial f_t}{\partial \hat{X}(t)} - K(t) \frac{\partial h_t}{\partial \hat{X}(t)} \right]' \\ & + \Gamma(t)Q(t)\Gamma'(t) + K(t)R(t)K'(t) \end{aligned} \quad (5-13)$$

with initial value  $P(t_0) = P_0$ .

Comparing (5-13) with (4-19), it is found that the two are similar. From (4-25), the approximate optimal matrix  $K(t)$  is

$$K(t) = P(t) \left[ \frac{\partial h_t}{\partial \hat{X}(t)} \right]' R^{-1}(t) \quad (5-14)$$

Similarly, according to (4-26), the differential equation followed by the filter error matrix of nonlinear unbiased and minimum-variance filters can be obtained:

$$\dot{P}(t) = \frac{\partial f_t}{\partial \hat{X}(t)} P(t) + P(t) \left[ \frac{\partial f_t}{\partial \hat{X}(t)} \right]' + \Gamma(t)Q(t)\Gamma'(t) - K(t)R(t)K'(t) \quad (5-15)$$

AD-A104 327

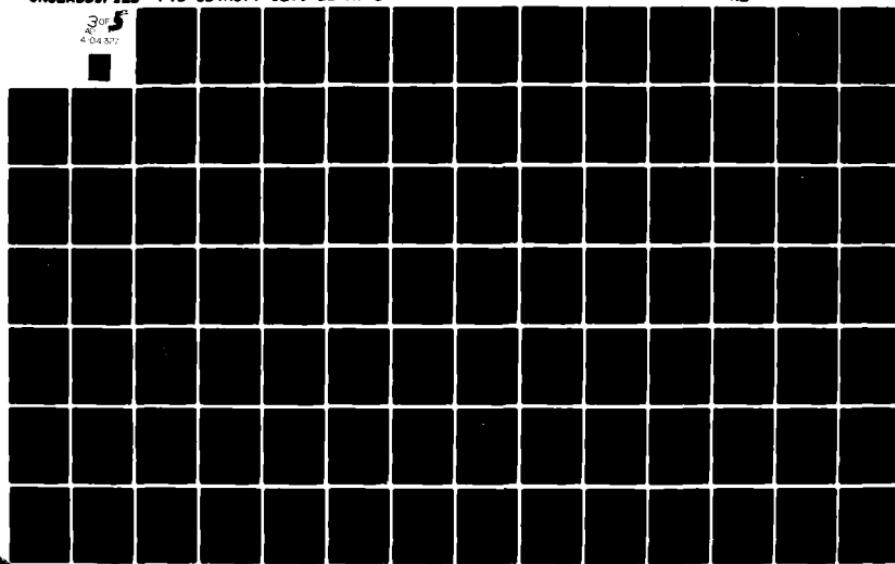
FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OH  
RECENT SELECTED PAPERS OF NORTHWESTERN POLYTECHNICAL UNIVERSITY--ETC(U)  
AUG 81  
FTD-ID(RS)T-0259-81-PT-1

F/0 20/4

NL

UNCLASSIFIED

3 of 5  
4 04307



Combining (5-3), (5-7), (5-8), (5-14), and (5-15), the complete equations of the nonlinear continuous-time systems' unbiased minimum-variance and suboptimal filters are obtained. In the above equations, if  $\alpha(t)$  and  $\beta(t)$  are 0 they become the first order suboptimal filters.

#### VI. CONCLUSIONS

This paper emphasized the correspondence relationship between the estimated system (original) and the filter (model). It first determined the structures and properties of linear discrete and continuous-time systems by considering the unbiased requirement. Furthermore, it determined the approximate compensation terms of the nonlinear discrete and continuous-time systems. Then based on the minimum-variance requirement, the optimal matrix  $K(\tau)$  for linear discrete and continuous-time systems is determined. Finally, the suboptimal matrix  $K(t)$  for unbiased and minimum-variance filters of nonlinear discrete and continuous-time systems is determined. For nonlinear systems, only second order approximation suboptimal filters are studied. It can in principle be extrapolated to the design of more advanced suboptimal filters.

## REFERENCES

- [1] Kalman, R.E., A New Approach to Linear Filtering and Prediction Problems, *Trans. ASME, J. Basic Eng.*, Vol. 82 D, 1960, pp. 34-45.
- [2] Kalman, R. E., and Bucy, R. S., New Results in Linear Filtering and Prediction Theory, *Trans. ASME, J. Basic Eng.*, Vol. 83 D, 1961, pp. 95-108.
- [3] Athans, M., and Tse, E., A Direct Derivation of the Optimal Linear Filter Using the Maximum Principle, *Trans. IEEE, Vol. AC-12*, 1967, pp. 690-698.
- [4] Athans, M., Wishner, R. P., and Bertolin, A., Suboptimal state Estimation for Continuous-Time Nonlinear Systems from Discrete Noisy Measurements, *Trans. IEEE, Vol. AC-13*, 1968, pp. 504-514.
- [5] Jazwinski, A. H., *Stochastic Processes and Filtering Theory*, Academic Press (1970), pp. 336.
- [6] Institute of Mathematics Chinese Academy of Sciences, Probability group. Mathematical method of discrete time system filtering, National Defense Industry Publishing House (1971), p. 187.

## Summary

# Solidification Characteristics of Superalloys under Non-equilibrium Condition

*Fu Hengzhi*

The development of unidirectional solidification and surface grain refinement process for turbine blades brings forth the problem of controlling crystallization structures of superalloys during solidification. An attempt is, therefore, made in the present paper to determine the effect of some metallurgical factors on the characteristics of solidification by means of thermal analysis and metallographic assessment. It was found that there exists a relationship between  $\tau_i/\tau_s$  (the ratio of holding time at the liquidus temperature to the total freezing time) and the macrostructure obtained in castings.

As the ratio  $\tau_i/\tau_s$  increases, successive solidification gradually becomes predominant in the freezing process, and, as a result, the grain growth changes from an equiaxed structure to an oriented columnar one. When  $\tau_i/\tau_s$  approaches unity, it is possible to grow a fully columnar structure.

The liquidus and solidus temperatures of Ni-Cr and Fe-Ni superalloys and the holding times at the critical temperatures during freezing were determined and are presented here. The effect of adding alloying elements to Ni-Cr alloys on  $\tau_i$  and  $\tau_s$  was also determined. With the exception of cobalt, the increase of which slightly enhances directional solidification, an increase of all the alloying elements, such as molybdenum (up to 22%), tungsten (up to 15%), niobium (up to 10%), aluminum (up to 9%) and titanium (up to 3%), leads to a decrease in the ratio  $\tau_i/\tau_s$ , hence hindering successive solidification.

Among all the alloying elements, aluminum and titanium are the most conspicuous in effectiveness. When the titanium content exceeds 1%, the oriented columnar growth is very rapidly made difficult to proceed.

$K$ , the solute distribution coefficient under non-equilibrium condition is also discussed. According to the expression  $\frac{G}{R} \geq \frac{mC_u(1-k)}{Dk}$ , which predicts the stability of directional solidification, the effective distribution coefficient, a modification of  $K$  (under equilibrium condition), is an important factor that affects the nature of grain growth, provided that other factors remain

unchanged.

The greater the deviation from equilibrium state during cooling, the greater will be the value of  $K$  and the smaller will be the difference between solute concentrations in solid and liquid phases. In such cases the build-up of solute at the solid-liquid interface will also drop, thus making constitutional supercooling, which hinder directional solidification, decrease. As to the quantitative determination of the solute distribution coefficient under non-equilibrium condition, it may depend on a number of metallurgical parameters.

SOLIDIFICATION CHARACTERISTICS  
OF SUPERALLOYS UNDER  
NON-EQUILIBRIUM CONDITIONS

Fu Hengzhi

Abstract

This paper is an attempt to analyze the solidification curves of alloys and to discuss the solidification characteristics of alloys under non-equilibrium conditions. Furthermore, it also includes a discussion of the effect of solute distribution coefficient on the crystallization process under non-equilibrium conditions. The effect of adding alloying elements for additional strength on the cooling curve and crystallization characteristics has also been investigated in a preliminary manner.

In order to improve the performance of airplane engines, it is common practice in other countries to incorporate a surface grain refinement process to control the crystallization structure on turbine blades in actual production lines. For example, Pratt & Whitney of the United States has already adopted a unidirectional solidification process to fabricate turbine blades. At present, crystalline and single crystalline turbine blades have been used in the U. S. and U. K. on TF-30-300, J-58 and other engines. More recently, an alloy prepared by the unidirectional solidification method (PWA-1422) will be used in the JT9D engine. The USSR also has used unidirectional solidified and single crystal-line alloys to fabricate turbine blades for AN-20 and AN-24 engines. The development of unidirectional solidification and surface grain refinement process for turbine blades brings forth the problem of controlling the crystallization structures of high

temperature superalloys during the solidification process. By controlling certain factors, it is possible to obtain either an oriented columnar or a uniaxial structure based on the mechanical properties required for each specific application.

This paper will attempt to discuss the effect of some metallurgical factors on the crystallization characteristics of high temperature alloys in order to gain some understanding regarding the conditions necessary for successive solidification.

#### I. Solidification Characteristics of Superalloys under Non-Equilibrium Conditions

Under an identical heat dissipation condition during casting, different alloy compositions tend to solidify into different structures, due to their own solidification characteristics. For the same alloy composition, if various degrees of excess heating exist, different crystalline structures can be obtained upon freezing in identical modes, even when the liquids were poured into those modes at the same temperature. This is due to the difference in liquid state. At the same pouring temperature, different crystal structures can be obtained with a low carbon heat resistant alloy containing 25% Cr by excessive heating of 0, 100, and 160 degrees C as well as by adding 0.3%  $T_i$  to the alloy for modification. With increasing degree of excessive heating, the columnar growth becomes thicker. When the 160°C excessive heating is followed by the addition of 0.5%  $T_i$ , the solidification time remains unchanged. However, the alloy structure changes from a thick columnar structure to a uniaxial structure (see Figure 1) [1]. It is noted, during an analysis of the alloy solidification process, that the total freezing time of the alloy remains basically the same, regardless of excessive heating and addition of other elements. But the holding time at the liquidus temperature ( $T_i$ ) appears quite different.

The ratio of  $\tau_1$  and the total freezing time ( $\tau_e$ ) is 0.51, 0.63, 0.73 and 0.37, respectively.

It is apparent that the increasing  $\tau_1/\tau_e$  values correspond to the columnar growth of the crystal based on comparison between the crystal structure and the  $\tau_1/\tau_e$  values mentioned above. When  $\tau_1/\tau_e$  is reduced from 0.73 to 0.37, despite the fact that the alloys have been heated to 1730°C, a uniaxial crystal structure along the casting object was obtained due to the addition of titanium.

During the solidification process, the holding time in the liquid phase represents the time period of constant temperature crystallization. The crystallization at constant temperature or near constant temperature indicates that solidification occurs successively. The freezing curves of the Ni 63, Cr 10, W 5, Mo 4, Al 5.5, Ti 2.5 alloy under unidirectional and regular solidification conditions fully demonstrated that the crystallization process takes place at temperature slightly below the liquidus temperature almost at a constant temperature until completion. The  $\tau_1/\tau_e$  ratio approaches 1 (see Figure 2). Under regular solidification conditions, however, the holding time of the alloy at the liquidus temperature is relatively short. The former case yields the typical successive unidirectionally oriented columnar structure while the latter brings about an uniaxial structure solidified almost simultaneously.

Therefore, the value of  $\tau_1/\tau_e$  reflects the characteristics of the crystalline structure of the alloy to some degree. The lower this ratio is, the higher the tendency to form the alloy via successive crystallization. The opposite then suggests that simultaneous solidification may have taken place. When  $\tau_1/\tau_e$  approaches 1, it means that it is possible to obtain a successively crystallized unidirectional columnar structure.

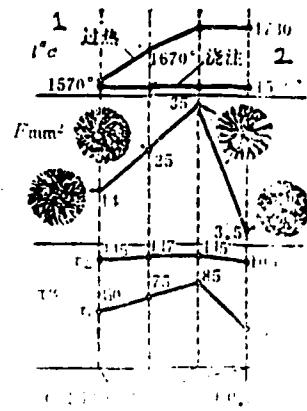


Figure 1. The effect of excessive heating on the solidification time and crystal structure.  
1) excessive heating; 2) pouring temperature.

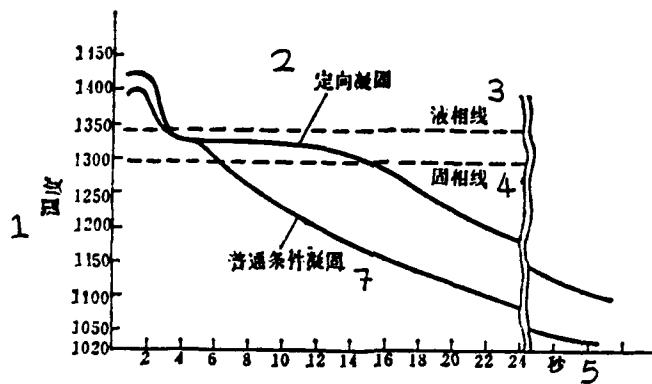


Figure 2. Solidification curves under regular and unidirectional conditions. 1) temperature;  
2) unidirectional solidification; 3) liquid phase;  
4) solid phase; 5) seconds; 7) regular solidification.

In nickel-based superalloys with the addition of various amounts of Al, Mo, W and Co, it is possible to obtain a curve as shown in Figure 3, which plots the values of  $\tau_1/\tau_e$  from the solidification curves and the fractions of columnar structure area in the cross-section of the cast object. The area occupied by the columnar structure increases with the  $\tau_1/\tau_e$  value. Therefore, it is possible to use the solidification curve of an alloy as one of the physicochemical parameters to characterize the properties of the crystal structure of the alloy. Table 1 shows the critical temperatures and solidifications times of certain Ni and Fe based superalloys. Their corresponding chemical compositions are listed in Table 2.

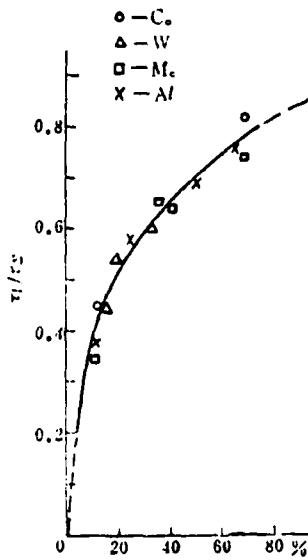


Figure 3. Relationship between  $\tau_1/\tau_e$  and columnar structure growth.

TABLE 1. CRITICAL TEMPERATURES (°C)  
AND SOLIDIFICATION TIMES (sec.) OF CERTAIN SUPERALLOYS

序号	$T_i$	$T_s$	$\Delta T$	$\tau_i$	$\tau_s$	$\tau_i/\tau_s$
1	1370	1295	75	30	135	~0.22
2	1345	1285	60	40	200	~0.2
3	3315	1245	90	0	124	0
4	1343	1289	74	54	375	~0.13
5	1380/1390	1350/1355	30/35			
6	1340	1300	40	130	245	~0.5
7	1370	1320	50	108	298	~0.4
8	1389	1314	75	122	192	0.6
9	1410	1785	125	168	325	0.5
10	1375	1340	35	35	95	0.37
11*	1330	1270	60	0	120	0
12*	1335	1290	40	30	150	~0.2

2 - code number

TABLE 2. CHEMICAL COMPOSITIONS OF THE ABOVE ALLOYS

序号	C	Ni	Cr	Co	Fe	Mo	W	Nb	Al	Ti
1	0.15	余 <sup>3</sup>	16		<8.0	4.0	5.5		2.0	2.0
2	0.15	余	15			4.5	7.7		4.5	1.5
3	<0.1	余	18			3.0	6.0		5.0	3.5
4	<0.1	余	16			15.0	3.0	2.0		
5	0.15	余	26		余		8.0			
6	<0.1	35	14		余		3.0		1.4	3.0
7	0.15	余	10		5.0	5.0	5.0		4.5	
8	<0.1	余	20			4.0	5.0	2.0		2.5
9	<0.1	20	16		余		3.0	1.2		
10	<0.12	38	15				3.0			1.3
11	0.15	余	11	6.0		4.0	5.0		2.5	2.5
12	0.15	余 <sup>1</sup>	10	10.0		4.0	5.0		2.5	2.5

2 - code number; 3 - balance.

It can be seen from the tables that the high Mo containing Ni-Cr alloy 4 has a relatively small  $\tau_t/\tau_e$  ratio. It is approximately equal to 1/2 and 1/3 of those of alloy 6 and 7, respectively. Comparing their structure under low magnification, it is apparent that the former is almost completely in uniaxial structure at the cross-section, while the cross-sections of the samples of the latter alloys contain considerable amounts of columnar grown crystals (see Figure 4).

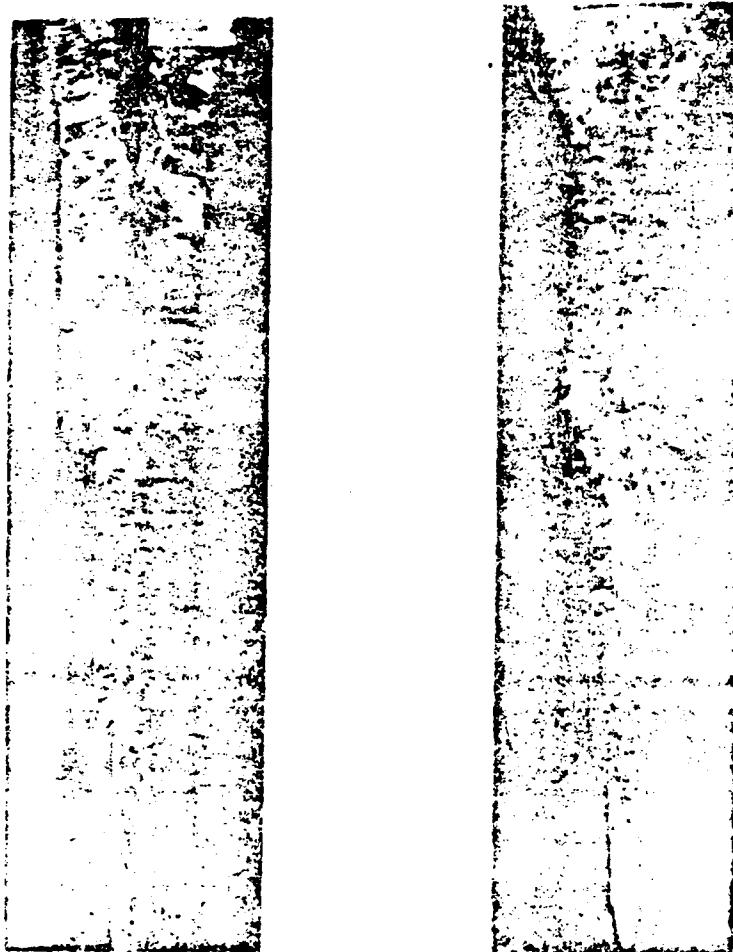


Figure 4. The low magnification structures of two superalloys.

For some alloys oversaturated with added strengthening elements, it is found that their solidification curves often show clear turning points between the solid and liquid phases. This is most probably due to the precipitation of some high melting point phase from the liquid in the form of metal carbides (MC) and intermetallic compounds. Based on preliminary investigation on alloys 11 and 12, the precipitation of MC and intermetallic compound  $r-r'$  occurs at the liquidus and solidus temperatures, respectively. The solidification curves of a Ni-Cr (16%) superalloy with the addition of 0-10% Nb and 0-22% Mo have been obtained. In both cases, an obvious turning point appears in the solidification curve. Based on metallographic analysis and phase identification, it was determined that  $Ni_3Nb$  and NbC were precipitated during the freezing process<sup>[2]</sup>.

Figure 5 shows the phase diagram of Fe-C alloys and the corresponding solidification kinetics scheme. The bottom plate shows the relation between the chemical composition of the alloy and the width of the oriented columnar structure zone. The shaded area represents the holding time of constant temperature crystallization<sup>[3]</sup>. It can be derived that the propagation of the columnar structure is approximately proportional to the  $\tau_i/\tau_e$  ratio obtained from the corresponding kinetic curve.

We can easily realize, from the kinetic solidification curves of various alloys, that alloys regardless of their solid structure (intermetallic or alloying compounds) tend to remain at constant current near the liquidus temperature once crystallization begins. This is consistent with the solidification curves of many superalloys obtained experimentally. Some people believe that the degree of freedom of an alloy within the range of temperature of solidification is not zero ( $f = R - n + l = 1$ ). Therefore, they consider it impossible for such constant temperature crystallization plateaus to exist. There should be another

degree of freedom "temperature" to be plotted. It is widely acknowledged that constant temperature crystallization plateaus only exist for pure metals and intermetallic compounds ( $f = 1 - 2 + 1 = 0$  for the former and  $f = 2 - 3 + 1 = 0$  for the latter). This discrepancy was brought about by applying an equilibrium condition to the non-equilibrium conditions encountered during casting. Under slow cooling equilibrium conditions, the latent heat liberated by the crystallization of the alloy makes it almost impossible to maintain a temperature plateau near the liquidus temperature. Assuming that the latent heat of crystallization is extremely large, and that the temperature of the alloy will rise to about the liquidus temperature, further crystallization can still not proceed. This is due to the fact that not only heat transfer property but also mass transfer characteristics must be satisfied at the fringe of the crystal before solidification can continue. For a binary alloy, under equilibrium condition, the composition of the initial solidified alloy and that for the final liquid phase differ from the original alloy composition by a factor  $K_C$  (where  $K_C$  is the solute distribution coefficient =  $C_S/C_O$  under equilibrium conditions) as shown in Figure 6. Assuming that there is no heat and mechanical convection in the liquid alloy, the movement of the solute then totally relies on diffusion. There must be a boundary layer as in thickness on the liquid side of the solid-liquid interphase in which the solute concentration far exceeds that of the bulk alloy.

The solute concentration distribution curve in liquid under equilibrium can be expressed as:

$$C_t = C_0 \left[ 1 + \frac{1 - K_C}{K_C} \exp\left(-\frac{R}{D}x\right) \right] \quad (1)$$

where

$R$  is the solidification rate, cm/sec

$D$  is the diffusion coefficient in liquid,  $\text{cm}^2/\text{sec}$ .

$X$  is the distance from the interphase, cm

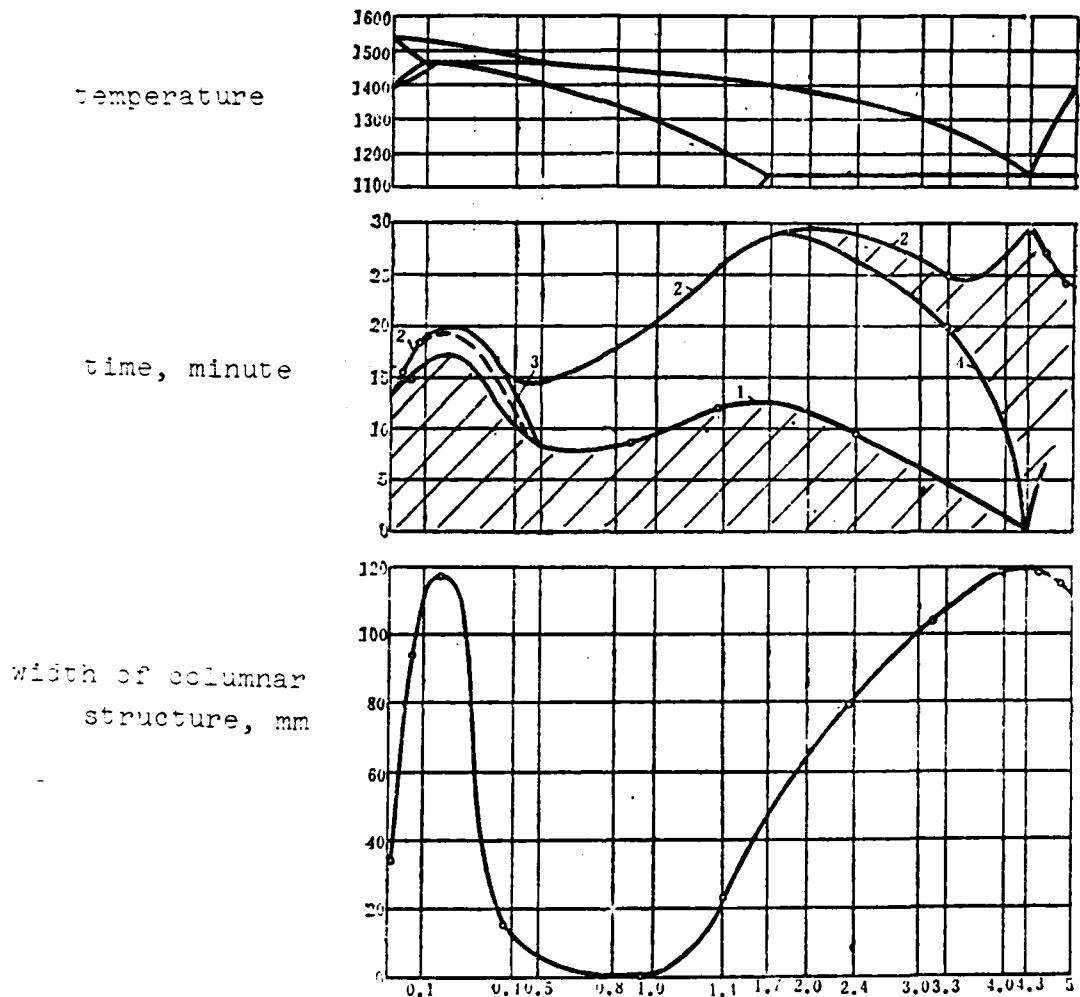


Figure 5. The chemical composition of Fe-C alloy versus solidification time and width of columnar structure. 1) liquid phase; 2) solid phase; 3) crystalline transition; 4)  $\nu'$  Fe precipitation.

The solute distribution function in solid is:

$$C_s = C_0 \left[ 1 - (1 - K_0) \exp \left( -\frac{K_0 R}{D} x \right) \right] \quad (2)$$

Based on boundary conditions, when  $x \rightarrow \infty$ ,  $C_1 = C_0$   
when  $x = 0$ ,  $C_1 = C_0 / K_0$

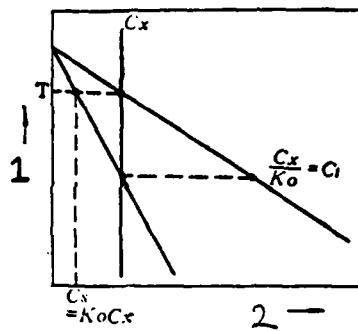


Figure 6. Binary phase diagram ( $K_o < 1$ ). 1) Temperature; 2) Concentration.

The corresponding concentration distribution curve is shown in Figure 7.

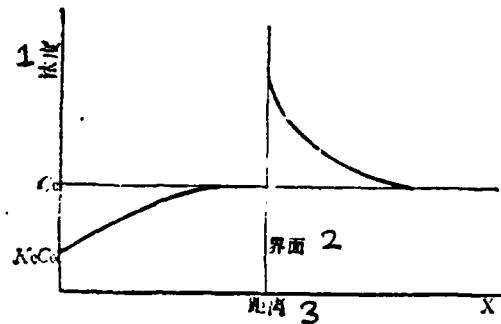


Figure 7. Solute distribution characteristic during solidification. 1) Concentration; 2) Interphase; 3) Distance.

The actual temperature of the alloy varies with the distance from the interphase, corresponding to solute concentration change discussed above. Figure 8 shows these distribution curves. The liquidus temperature of the alloy can be written as:

$$T_l = T_0 - m C_0 \left[ 1 + \frac{1 - K_0}{K_0} \exp \left( - \frac{R}{D} \right) \right] \quad (3)$$

The actual temperature of the liquid  $T$  during cooling, however, is a function of the temperature gradient in the liquid and the distance away from the interphase:

$$T = T_0 - m (C_0 / K_0) + G_0 \quad (4)$$

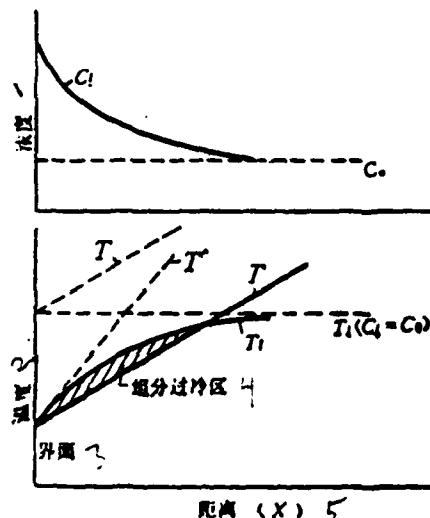


Figure 8. The variation in liquid concentration and temperature during solidification.

1 - concentration; 2 - temperature; 3 - interphase;  
4 - cold composition zone; 5 - distance (x).

where  $m$  is slope of the liquid phase curve

$T_0$  is the melting point of the pure solute metal,  $^{\circ}\text{C}_0$   
and  $G$  is the temperature gradient,  $^{\circ}\text{C}/\text{cm}$ .

In Figure 8, if we assume that the concentrated solute boundary layer does not exist ( $C_1 \rightarrow C_0$ ), then the liquidus temperature of the alloy  $T_1$  remains constant. It intersects with the actual temperature distribution curve at the interphase. When cooling continues, successive crystallization will then be the process of solidification. Since the concentrated boundary layer of the solute does exist, the liquidus temperature also decreases. Under this condition, external cooling to reduce the liquid temperature is necessary to continue the solidification process. Even though the crystallization continues, it is easy to notice that successive solidification has been interrupted due to the formation of new nuclei in the cold region away from the interphase. In order to avoid this effect, the slope of the temperature distribution curve of the alloy's liquidus temperature at the interphase must be smaller than that of the actual temperature distribution curve of the liquid at the interphase:

$$\frac{dT}{dx} \geq \left( \frac{dT_1}{dx} \right)_{...}$$

It is possible to derive from (3) and (4) that the condition necessary to maintain successive crystallization is:

$$\frac{G}{R} > \frac{mC_0(1-K_0)}{DK_0}$$

In order to obtain successive solidification and to avoid a super cold region, it seems that we should maximize the

solidification rate. Obviously, it is advantageous to increase the temperature gradient in the liquid without any doubt. However, the minimization of the solidification rate to increase the G/R ratio for successive crystallization deserves further consideration. Extremely slow solidification rate not only significantly lengthens production time but also leads to localized preferential precipitation, drastically lowering the characteristics of the casting object. Research and development on unidirectional crystal growth indicated that drastic reduction in solidification rate only lowered the distribution coefficient, leading to localized uniaxial crystal formation. [4,5].

In actual casting of objects, the cast part is usually cooled at a significant rate with a temperature gradient in the alloy which is unlike the equilibrium case. Whether the latent heat released from crystallization can maintain the liquidus temperature depends on the condition under which non-equilibrium crystallization progresses. If the solid composition is fairly close to that of the liquid phase, crystallization can still take place under isothermal conditions. The composition of the liquid phase remains unchanged, while the solid phase precipitates out of solution, which means the continuous crystallization isothermally. Under such conditions, the solid and the liquid become one unit ( $f = 1 - 2 + 1 + 0$ ). Thermodynamically, it is possible for alloys to solidify in nearly the same composition as their corresponding liquid phase deviating from the compositions of liquid and solid phases under equilibrium. For example, as shown in Figure 9, alloy X solidifies from the liquid phase and begins crystallization at temperature  $T_1$ . The equilibrium solid state composition is  $C_s$ . This is because the composition  $C_s$  has the lowest energy in the solid state at temperature  $T_1$ . Actually, as long as nuclei are being produced in the alloy, precipitation of the solid phase under non-equilibrium conditions remains possible for solid state compositions to the left of  $C_s$  in Figure 9. This

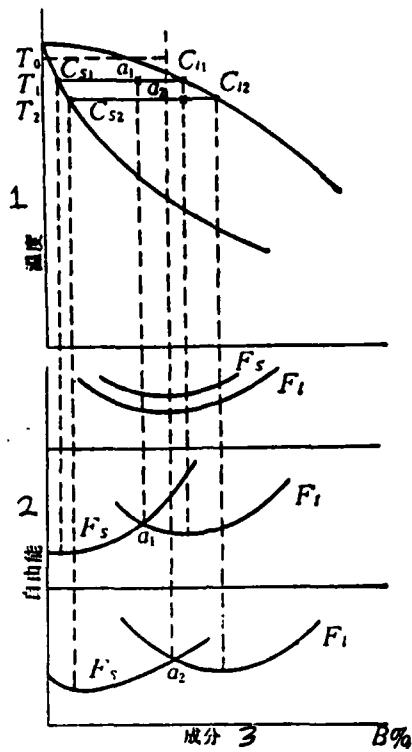


Figure 9. Alloy crystallization and its corresponding change in free energy: 1) temperature; 2) free energy; 3) composition.

is because to the left of point a, the free energy for the liquid phase is always higher than that of the solid state, regardless of composition. Similarly, as the temperature decreases to  $T_2$  or  $T_3$ , solid states with compositions to the left of point a can precipitate. The composition of the corresponding liquid phase depends upon the ratio of the two phases and the non-equilibrium distribution coefficient. It is closer to that of the liquid alloys under ordinary conditions than under equilibrium conditions. As a matter of fact, non-preferential crystallization can be accomplished and isothermal successive solidification obtained once an alloy is cooled at a fixed rate to the low temperature at which the free energy of the precipitated solid state is equal to or less than that of the original liquid alloy.

Therefore, under non-equilibrium conditions, the effective solute distribution coefficient  $K$  is greater than its equilibrium value  $K_0$  due to various cooling conditions, adjustments in chemical compositions and the variation of the physicochemical deviation from equilibrium. As the value of  $K$  becomes greater, the concentration dam and concentrate gradient become smaller. The parameter representing the difference between solute concentration in the initial solid phase and the residual liquid phase then becomes  $(\frac{1-K}{K})$  non-equilibrium  $< (\frac{1-K_0}{K_0})$  equilibrium. Under these conditions, the liquid phase curve which reflects the variation of solute concentration should be leveling off. Figure 10 shows various liquid phase curves based on different  $K$  values. The parameters chosen are:

$T = 1454^{\circ}\text{C}$  (melting point of pure Ni)

$m = \tan 30^{\circ}\text{C}$  (Average slope of the liquid phase curve of a Ni alloy containing Ti, Al and Nb)

$C_0 = 10\%$

$G = 50^{\circ}\text{C/cm}$

$R = 0.02 \text{ cm/sec}$  (unidirectional solidification rate)

$D = 10 \text{ cm}^2/\text{day}$ <sup>[6]</sup> (diffusion coefficient of viscous liquid metal)

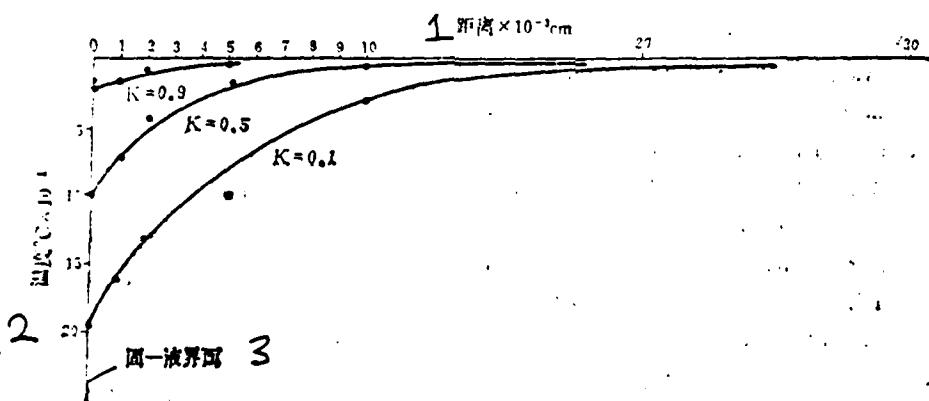


Figure 10: Liquid phase curves of an alloy at different  $K$  values:  
 1) distance  $\times 10^{-3}$  cm; 2) temperature  $\text{^{\circ}C} \times 10^{-1}$ ; 3) solid-liquid interphase.

From the distribution curves of the liquid phase shown in the figure, it is clearly demonstrated that the larger the  $K$  value is, the easier successive solidification becomes. Based on calculation, the concentration dam is practically non-existent at the interphase when  $K$  value is larger than 0.7 - 0.8. Crystallization process becomes almost completely non-preferential. In other words, the composition variation between the solid and liquid states becomes less under non-equilibrium conditions. On the other hand, when the solute distribution coefficient decreases due to certain conditions (e.g., grain refinement processing), the solidification of the alloy becomes a simultaneous process favoring the formation of small uniaxial structures. Figure 11 shows the variations of the Cr distribution coefficient  $K$  and Preferential Precipitation Coefficient  $J$  ( $C_{\max}/C_{\min}$ ) of the cross-section of cast Cr-Mo steel in the columnar as well as the uniaxial zones<sup>[7]</sup>. It can be seen that the Cr distribution coefficient decreases from 0.9 to 0.7 as the columnar structure changes over to the uniaxial structure away from the surface of the cast ingot. Figure 12 shows the relation between the alloy grain size and the casting parameter  $(\frac{-m_{CO}(1-K)}{K})$ <sup>[8]</sup> for a Ni alloy with the same atomic percent of Group II elements. Those elements less soluble in nickel become surface active elements which drastically change the characteristics of nickel-based alloys in the liquid phase. They tend to increase  $\frac{-m_{CO}(1-K)}{K}$ , leading to the significant changes in the crystal structures.

In summary, the effect of solute distribution coefficient on the characteristics of the crystal during alloy solidification is very significant. Under non-equilibrium conditions, to change or to modify this coefficient can help us to control the crystallization process to a certain degree of importance in order to obtain the desired structure. Through the measurement and analysis of the alloy solidification curve for determination of the  $\tau_1/\tau_\epsilon$  value, we are also aided in qualitatively understanding the variation of the solute distribution coefficient.

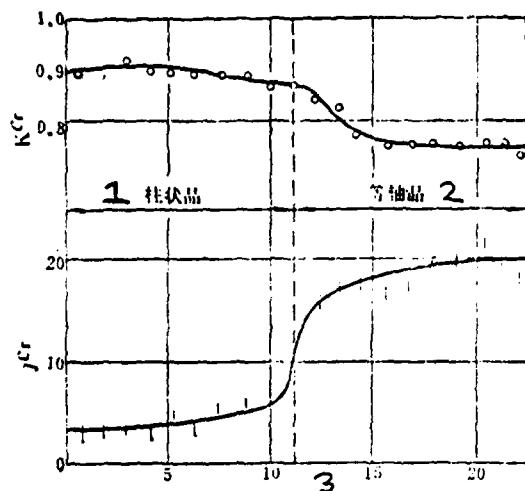


Figure 11. Relation of  $K^{Cr}$  and  $J^{Cr}$  with distance to surface of a Cr-Mo steel: 1) columnar crystal; 2) uniaxial crystal; 3) distance from the surface of cast ingot, cm.

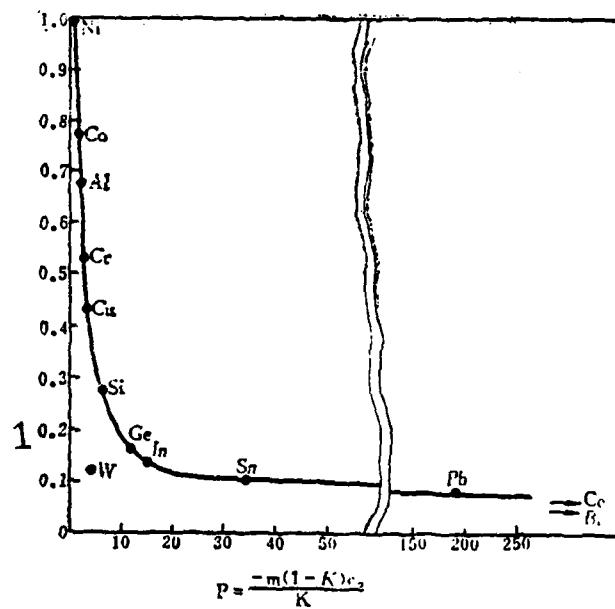


Figure 12. Relation between grain size and casting parameter: 1) relative grain size.

Along with the non-preferential crystallization process, successive solidification and the plateau in the liquid phase line in the solidification line indicate that solidification progresses under non-equilibrium conditions. The longer the holding time at the liquidus temperature, the larger the corresponding inception period of non-equilibrium solidification becomes. Therefore, the value of  $\tau_1/\tau_\epsilon$  serves as an indicator to show the degree of deviation away from equilibrium for the freezing alloy. Smaller values of  $\tau_1/\tau_\epsilon$  usually mean higher thermodynamic stability and vice versa. As the successive crystallization continues, the composition of the liquid phase of the alloy deviates more and more away from its equilibrium value and the instability of crystallization becomes more serious. Once this instability can no longer be maintained, it is then impossible to crystallize at the liquidus temperature of the alloy. The plateau in the solidification curve for the near non-preferential crystallization process disappears. It becomes a declining curve indicating one degree of freedom. The fact that usually the exterior of a cast object has a columnar structure while its interior shows mostly a uniaxial structure may reflect the above situation.

Successive crystallization under non-equilibrium conditions produces solidified alloys with compositions different from the ones obtained under equilibrium. Even under extremely non-equilibrium and near non-preferential crystallization conditions, the chemical compositions at the crystal axes and crystal spacings are different with larger concentrations of elements which lower the melting point of alloys at the latter. For example, Al, Ti, Mo, Nb and C at higher than their equilibrium concentrations can be precipitated temporarily during non-equilibrium solidification. This is quite unusual, since the crystal axes which solidify first should be enriched by those lower melting point elements. This preferential precipitation effect becomes less pronounced along grain boundaries and between grains for certain superalloys. Figure 13 shows the

micrographs of a Ni-Cr-Mo-W-Nb alloy from its columnar and uniaxial zones. Using x-ray diffraction, it has been determined that the precipitated phases are Laves  $[Ni_2(Nb,Mo)]$  and  $M_6C[Cr_3Mo_3C]$ . It was observed that these precipitates are more concentratedly located at the grain boundaries and between the two crystal structures for the uniaxial structural sample. The precipitation of  $M_6C$  and Laves at the grain boundaries and along the side branches becomes rare in samples obtained with a columnar structure, showing an unusual depletion effect. Similar results were reported in reference [4]. Electronic probes have been used to determine the Ti and W contents for U-700 and Mar-M200 alloys along the main axes and the side branches. The preferential precipitation ratios for Ti are 1.19, 1.21 and 1.06 for areas along the main branches, near the side branches and an average zone, respectively. For W, the ratios are 0.62, 0.77 and 0.91. In other words, alloy composition variation is larger in those areas when the ratio deviates from 1. Especially, in the case of W, the area between the uniaxial and columnar zones appears to have a higher preferential precipitation ratio than the main branches (deviation from 1 is more severe). This implies the effect of little composition variation during solidification due to the large distribution coefficient for columnar crystals. Similarly, due to the fact that the distribution coefficient is always larger for unidirectional columnar crystal growth than that for uniaxial crystallization, the precipitation of carbon compound along the grain boundary and the side branches becomes far less than the case of uniaxial crystallization. For Udinet-700 alloy, carbon precipitate in the oriented columnar zone is far less concentrated than in the uniaxial structure. Picarcy<sup>[9]</sup> studied the characteristics of unidirectional solidification of MM-200, In-100, B-1900 and TRW-1900 superalloys. It was found that during conventional casting, the precipitation of  $r'$  is extremely non-uniform. The concentration is much higher in grain boundaries and in between branches. In unidirectional solidification, however, the distribution of  $r'$  is more uniform with significant precipitation along the axes. This



a) 取自等轴晶区



b) 取自柱状晶区

Figure 13. Micrographs of a Ni-Cr-Mo-W-Nb alloy:  
1. a) from uniaxial structure; 2. b) from oriented  
columnar structure.

is more apparent for TRW-1900 and B-1900 alloys. For In-100 and B-1900 alloys, during unidirectional solidification, it was found that only  $r'$  was detected structurally in the  $r-r'$  crystal. This indicates that the  $r$  structure has been dissolved during the solidification step. This also offers an explanation to the fact that under unidirectional solidification, the residual liquid phase, unlike under equilibrium, provides the solid dissolution conditions.

## II. THE EFFECT OF ALLOYING ELEMENT ON THE CHARACTERISTICS OF NON-EQUILIBRIUM SOLIDIFICATION

The nickel content in an iron-based alloy (with 0.1% C and 20% Cr) has great influence on the characteristics of non-equilibrium solidification. Furthermore, its effect on the physical parameters of the alloy is significant. From the solidification curve as well as based on the structural analysis of the cross-section, the total solidification time and the holding time at liquidus temperature are at maximum values when

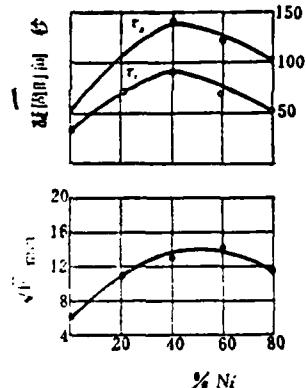


Figure 14. 1) solidification time.

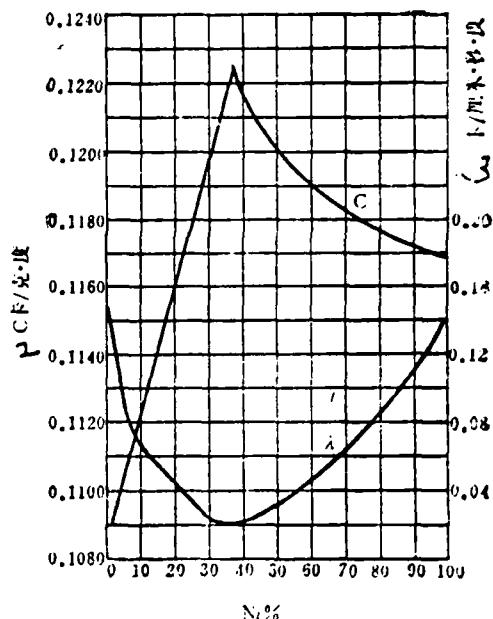


Figure 15  
2)  $k$  cal/ $\text{cm} \cdot \text{sec} \cdot {}^{\circ}\text{C}$   
3)  $k$  cal/ $\text{cm} \cdot \text{sec} \cdot {}^{\circ}\text{C}$

the nickel content is at 40-50% (see Figure 14.) Measurements of thermal conductivity coefficient of nickel-based alloy with respect to the nickel content reveal that a 40-50% Ni iron-based alloy has the smallest thermal conductivity coefficient and the maximum heat capacity as shown in Figure 15<sup>[10]</sup>. With increasing Ni concentration, thermal conductivity increases and the heat capacity drastically decreases. It is natural to interpret the fact that the alloy solidification time decreases once the nickel content exceeds 50% based on the thermophysical parameters of the alloy.

It can be observed from the structure of the cast alloy that the successive or simultaneous solidification characteristics do not vary noticeably with increasing nickel content in the alloy. However, due to the lengthening of the total solidification time, the grain size becomes significantly larger. The relation between the nickel content and the grain size corresponds very well with the dependence of nickel content on the solidification time.

In a study of the Ni-Cr-Mo-W-Nb superalloy series, an

TABLE 3. EFFECT OF Mo, W, Nb, Co ON THE SOLIDIFICATION TIME OF ALLOYS

ALLOY	$\tau_1$	$\tau_\epsilon$	$\tau_1/\tau_\epsilon$	ALLOY	$\tau_1$	$\tau_\epsilon$	$\tau_1/\tau_\epsilon$
金	$\tau_1$	$\tau_\epsilon$	$\tau_1/\tau_\epsilon$	金	$\tau_1$	$\tau_\epsilon$	$\tau_1/\tau_\epsilon$
Ni80Cr18	49	102	0.48	Ni80Cr18	49	102	0.48
+ 2%Nb	42	100	0.42	+ 3%W	45	66	0.47
+ 5%Nb	35	94	0.37	+ 5%W	45	90	0.5
+10%Nb	30	85	0.35	+15%W	40	87	0.45
+ 3%Mo	45	98	0.47	+ 5%Co	46	100	0.46
+10%Mo	34	91	0.38	+10%Co	50	91	0.55
+15%Mo	30	90	0.33	+15%Co	48	88	0.57
+22%Mo	24	88	0.27				

Co and Ni can form a continuous solid solution. Co is the only element capable of raising the melting point of the alloy with tungsten. Cobalt is also the only one to increase the  $\tau_1/\tau_\epsilon$  value although only to a very limited extent. From the solidification time listed in Table 3, the addition of cobalt to the alloy does not affect the holding time at the liquidus temperature, but it quite significantly decreases the total solidification time  $\tau_1$ . In actual poured step samples, the percentage of area covered by columnar crystals increases drastically with the cobalt content in the alloy. Among the Ni-Cr-Mo-Nb alloys with various amounts of cobalt, columnar crystals grow far better than in alloys without any cobalt present. Especially in alloys containing 15-20% cobalt, columnar crystals propagate through the cross-section of the sample.

Elements such as Al and Ti are currently used as the major

precipitation strengthening additives. Their effect on the  $\tau_1/\tau_\epsilon$  value appears to be more significant than that of Mo, W and Nb.  $\tau_1/\tau_\epsilon$  value is drastically reduced, which practically blocks the successive crystallization of the alloy. Unlike those high melting point elements, such as Mo, W and Mo, which decrease the  $\tau_1/\tau_\epsilon$  value by reducing the holding time of the alloy at the liquidus temperature during solidification, Al and Ti can significantly increase the total solidification time to reduce the  $\tau_1/\tau_\epsilon$  value. At the present time, the aluminum content in the superalloy series is less than 6%. Within this range, the amount of aluminum added basically does not vary the  $\tau_1/\tau_\epsilon$  by much. Therefore, it is negligible. However, the effect of titanium on the solidification characteristics is worth noticing. Based on the experimental results discussed above, titanium at more than 1% can drastically promote the formation of uniaxial crystals, largely affecting the solidification characteristics of the alloy. Therefore, titanium is the element which causes composition differences between the liquid and the solid states. The temperature range within which an alloy crystallizes also affects the solidification characteristics. Titanium is the most influential element in widening the range in Ni-Cr based superalloys. Based on report [11], less than 1% Ti would significantly decrease the alloy's solidification heat, reducing the solidification time. More than 1% titanium, on the contrary, would produce an exothermic reaction, promoting the formation of uniaxial crystals and other defects.

Based on the experimental results by Cook, et al<sup>[12]</sup>, the addition of Co, Cr, Mo and V to nickel-based alloys does not cause an apparent loosening effect. Only titanium can prompt preferential precipitation and loosening. On the commonly used In-100 superalloy, due to the 5.2% high titanium content, the blade produced by casting has been found to be seriously affected by preferential precipitation and loosening effects. In maintaining the same strength, Al content was increased to 6.7% and Ti content was decreased to 2.5% in order to alleviate the problem. An alloy

completely free of titanium which contains 6.3% Al, 11% W and 1.5-3% Nb-Ta alloy was formulated to cast parts without defects mentioned above.

In regard to unidirectional casting, those elements which cause a decrease in  $\tau_l/\tau_e$  and the solute distribution coefficient are deleterious to successive crystallization. Special precautionary measures must be taken in the design and application of unidirectional solidification of superalloys. With the rising demand for high temperature alloy performance, focus should be placed on the precipitation of high melting point compounds during solidification which may serve as nuclei for crystallization. They usually begin to appear during the early stage of solidification at temperatures slightly lower than the liquidus temperature. These compound particles exist in front of the solid phase prompting simultaneous crystallization.

### III. CONCLUSIONS

In the research and development of unidirectional solidification, the study of the variation of solute distribution coefficient during solidification based on the heat transfer of the cast object, profoundly discussing the effect of the solidification technology and the physical parameters of the alloy on the solute distribution coefficient, may be very important. The effect of solute distribution coefficient K on the metallurgical defects such as preferential precipitation, loosening and thermal cracking during the casting of blades is significant. More detailed analysis is certainly mandatory. This paper attempts to discuss the relationship between the solidification kinetics and solute distribution coefficient through the analysis of the solidification curve. The work is preliminary and crude. We welcome further discussion and comments in order to stimulate wider interest in the research in this area.

#### REFERENCES

[1] Н.Г.Гиршович, Ю.А.Нехеидзи, Затвердевание отливок. Сборник "Затвердевание Металлов". Машгиз (1958), стр. 39-90.

[2] Fu Hengzhi, The effects of molybdenum and niobium on the structure and properties of nickel-based castings of heat resisting alloys. Jinshuxuebao [metallurgical journal], Vol 8, No. 2(1965), pp. 212-220

[3] Б.Б.Гулжев. Литейные процессы. Машгиз (1961), стр. 182, 343.

[4] Giamei, A. F. and Kear, B. H., On the nature of freckles in Nickel Base Superalloys, Met. Trans. Vol. 1, No. 8 (1970), pp. 2185-2192.

[5] Coply, S. M. Giamei, A. F., Johnson, S. M. and Hornbecker, M. F.. The origin of freckles in unidirectional solidified castings, Met. Trans., Vol. 1, No. 8 (1970), pp. 2193-2204.

[6] Frenkel, J. Kinetic theory of Liquids. Oxford University press. (1946), pp. 201.

[7] Part 4 of translated text on directional solidification and mold casting. Shanghai Institute of Scientific and Technological Information (1977), pp 61-69

[8] Tarshis, T. A., Walker, J. L., and Rutter, J. W., Experiments on the solidification structure of alloy castings, Met Trans., Vol. 2, No. 9, (1971), pp. 2589-2597.

[9] Piearcy, B. J. and Terkelsen, B. E., The effect of unidirectional solidification on the properties of cast nickel-base superalloys. TMS-AIME, Vol. 239 No. 8 (1967), pp. 1143-1150.

[10] Fu Hengzhi, Investigation of the effects of composition and vacuum on the structure and properties of heat resisting alloys. Dissertation for candidate of Tech. Sci. (1962).

[11] Ю.А.Нехеидзи, Литейные свойства Жаропрочных сплавов. Машгиз (1963), стр. 76.

[12] Cook, R. M., and Guthrie, A. M. Factors affecting the foundry characteristics of Nickel-rich high-temperature alloys. Proceedings of First World Conference (1966).

## Summary

# The Plastic Deformation, Micro-Crack Initiation, and Fatigue Crack Initiation Life of 30CrMnSiNi2A High Strength Martensite Steel

*Zheng Xiulin, Qiao Shengru, and  
three 1978 graduates*

The significance of fatigue crack initiation life (FCIL) was discussed in references[1][2]. The FCIL is closely related to the mechanism governing the initiation of fatigue crack, which, however, is not yet quite clearly understood for high-strength Martensite steel. In order to clear up the mechanism governing the initiation of fatigue crack, it is necessary to study the mechanism governing plastic deformation, which, however, is so far incompletely understood.

Systemic research in the correlation between the fatigue crack initiation, the constitution, and the structure morphology of steels is also lacking. Reference[6] points out that the retained Austenite in the 30CrMnSiNi2A high-strength Martensite steel lowers the yield stress and shortens the FCIL. Evidently the amount of retained Austenite and its distribution have a great influence on plastic deformation and fatigue crack initiation.

Reference[7] points out that the residual compressive stress in the surface layer not only prolongs the FCIL but also decreases the fatigue crack propagation rate, so that the total fatigue life is also prolonged. However, Reference[7] did not deal with the effect of local prestrain at notch roots on the FCIL.

This paper presents the investigation of the structure, plastic deformation, micro-crack initiation, and the FCIL of 30CrMnSiNi2A high-strength Martensite steel by means of optical microscope, electron microscope and impact fatigue testing machine.

Test results show that lathy Martensite is the predominating structure of 30CrMnSiNi2A steel when oil-quenched and martempered. In addition, some retained Austenite of small amount was found in all tests.

It was found that, on the polished surfaces of test pieces, the so called "deformation reliefs" appear in the plastic zone around the tip of the fatigue

crack. As the crack propagates, it leaves behind itself the deformation reliefs on both sides of the seam. After polishing off the deformation reliefs and then etching the polished surface, Martensite laths can be found lying right under the deformation reliefs. Both the orientation and the size of the deformation reliefs correspond exactly with those of the Martensite laths. On the basis of these observations, it is reasonable to suggest that the appearance of the deformation relief is due to the sliding of Martensite laths, which is an important mode of plastic deformation of the lathy Martensite structure. When the amount of slip exceeds a certain limit, the micro-crack is initiated along the boundaries between Martensite laths.

The retained Austenite, distributed in the form of thin layers on the boundaries of Martensite laths, facilitates sliding, thus lowering the yield stress, promoting the crack initiation, and shortening the FCIL of the 30Cr-MnSiNi2A steel. The more the amount of retained Austenite, the larger is this effect.

The local prestrain at notch roots has a great effect on the FCIL. Prestrain in the positive direction (i. e., in the same direction as that of the cyclic stress) prolongs the FCIL, while prestrain in the negative direction shortens it.

The Plastic Deformation, Micro-Crack Initiation  
and Fatigue Crack Initiation Life of 30 CrMnSiNi2A  
High Strength Martensite Steel

Zheng Xiulin, Qian Shengru, Zhang Jienguo, Lou Bin  
and Gao Pushien

ABSTRACT

This paper presents the investigation of the structure, plastic deformation, micro-crack initiation, and the fatigue crack initiation life (FCIL) of 30CrMnSiNi2A high-strength Martensite steel by means of optical microscope, electron microscope and impact fatigue testing machines. Test results show that lathy Martensite is the predominating structure of 30CrMnSiNi2A steel when oil-quenched and martempered. It was found that, on the polished surfaces of test pieces, the so-called "deformation reliefs" appear in the plastic zone around the tip of the fatigue crack. Both the orientation and the size of the deformation reliefs correspond exactly with those of the Martensite laths. On the basis of these observations, it is reasonable to suggest that the appearance of the deformation relief is due to the sliding of Martensite laths, which is an important mode of plastic deformation of the lathy Martensite structure. When the amount of slip exceeds a certain limit, the micro-crack is initiated along the boundaries between martensite laths. The retained Austenite, distributed in the form of thin layers on the boundaries of Martensite laths, facilitates sliding, thus lowers the yield stress, promotes the crack initiation, and shortens the FCIL of the 30CrMnSiNi2A Steel. The local prestrain at notch roots has a large effect on the FCIL. Prestrain in the positive direction prolongs the FCIL, while prestrain in the negative direction shortens it.

## INTRODUCTION

The significance of fatigue crack initiation life (FCIL) was discussed in Refs. [1] and [2]. Factors which affect the FCIL were also discussed in Ref. [2]. The FCIL is closely related to the mechanism governing the initiation of fatigue crack.

Ref. [3] proposed a fatigue crack initiation model, which is based on the research results on steel fatigue crack initiated by the cracking of clamped impurities from the boundary of base material under alternating stress. Fatigue cracks can also be promoted through boundary sliding, or through the cracking of the crystal boundary [4]. However, the mechanism governing the initiation of fatigue crack is not yet clearly understood for high-strength Martensite steel. In order to clear up the mechanism governing the initiation of fatigue cracks, it is necessary to study the mechanism governing plastic deformation. Plastic deformation of the Martensite steel has been discussed in Ref. [5] which, however, is not satisfactory. Apparently, the mechanism governing plastic deformation of Martensite steel is so far incompletely understood and further study is necessary.

Systematic research on the correlation between fatigue crack initiation, the constitution, and the structure morphology of steels is also lacking. Reference [6] points out that the retained Austenite in the 30CrMnSiNi2A high-strength Martensite steel lowers the yield stress and shortens the FCIL. Evidently the amount of retained Austenite and its distribution have a great influence on plastic deformation and fatigue crack initiation.

It is well-known that fatigue strength can be enhanced by the residual compressive stress in the surface layer to prolong fatigue life. On the contrary, the residual pulling stress will downgrade the fatigue strength and shorten the fatigue life. Reference [7] points out that the residual compressive stress not only prolongs the FCIL but also decreases the fatigue crack propagation rate, so that the total fatigue life is also prolonged. However, the effect of local prestrain at notch roots on the FCIL has not been dealt with yet.

Under different heat treatment, the composition of 30CrMnSiNi2A steel may be different. In particular, the amount of the retained Austenite is apparently different. The aim of our study is to provide various compositions of the 30CrMnSiNi2A steel with various heat treatment systems, so that the correlation between the plastic deformation, micro-crack initiation, and the constituents of the high-strength steel can be observed, and the effect of heat treatment, constituents, and prestrain on the FCIL can be determined.

#### SAMPLE PREPARATION AND TESTING APPROACH

The chemical constituents of the 30CrMnSiNi2A steel used in our test are: 0.30%C, 1.09%Cr, 1.16%Mn, 1.09%Si, 1.51%Ni, 0.09%Cu, 0.04%S, 0.011%P.

Both the shape and the size of the sample are shown in Figure 1.

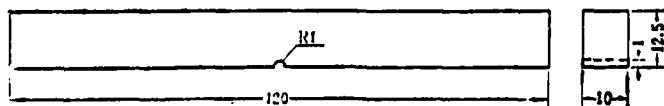


Figure 1.

The flow chart of the fabrication process for the sample is given below:

Forging → prepare for heat treatment  
(650°C annealing 3 hours) → cutting  
→ final heat treatment → grinding  
→ notch opening.

The final heat treatment system for the sample listed in Table I. The notch of the sample is cleaved with an optical grinding machine.

The prestrain is applied in two cases: 1) the sample is ground, cleaved, open notch and prestrained under martempering conditions (900°C Austenite process, after 60 minutes of 230°C tempering, cooling to room temperature). Then the sample is heated to 275°C and annealed for 3 hours. The prestrain method is as shown in Figure 2a. The load is  $P = 1250$  kg, nominal stress is  $\sigma = 120$  kg/mm<sup>2</sup>. Since the loading direction of prestrain is opposite to that of fatigue, it is called negative prestrain; 2) the sample, after direct tempering (900°C Austenite process, oil-quenched), is annealed for 3 hours at 200°C. The prestrain is imposed after grinding, cleavage, and notch opening. The method of imposing prestrain is shown in Figure 2b. The load is  $P = 1400$  Kg, and the nominal stress  $\sigma = 135$  kg/mm<sup>2</sup>. Since the loading direction of the prestrain is same as for fatigue, it is called positive prestrain.

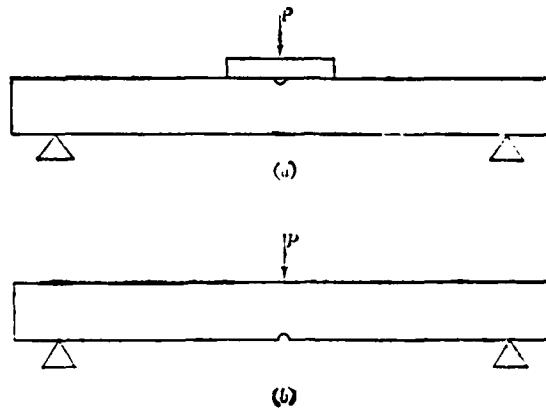


Figure 2. Loading method for prestrain  
 a) Negative prestrain; b) positive restraint

Fatigue testing is performed on a DSW0-150 type impact fatigue testing machine. Impact energy  $W_I = 207.5 \text{ kg-mm}$ . Maximum cyclic load  $P_m$  is calculated following the formula:

$$P_m = \left( \frac{2W_I}{C} \right)^{1/2} \quad (1)$$

The C in Eq. (1) is the softness coefficient of the notch sample, which can be calculated based on the formula given in Ref. [10]. Calculated result:  $P_m = 798 \text{ kg}$ ; experimental measurement  $P_m = 834 \text{ kg}$ . The maximum cyclical stress at the notch roots is  $\sigma = 155 \text{ kg/mm}^2$ . The minimum cyclical load is zero; thus the minimum cyclical stress is also zero, and stress ratio  $R = 0$ . The loading frequency is  $f = 225/\text{minute}$ .

Both sides of the sample are examined in order to observe cracks and surface condition. The length of the crack is measured by means of a microscope with an accuracy of  $0.001/\text{mm}$ . When any side of the sample shows a crack  $0.2 \text{ mm}$  in length, the number of stress cycles experienced is defined as the fatigue crack initiation life [2].

The variation of the surface condition is observed by means of an optical microscope, while the structure of steel is observed by means of an optical microscope and an electron microscope.

#### TEST RESULTS

##### Structure Morphology

The structure of 30CrMnSiNi2A steel, after 20 minutes of the Austenite process at 900°C, oil-quenching, 200°C annealing for 3 hours, is primarily a lathy Martensite (Figure 3); but a small amount of Austenite remains.

The structure of the 30CrMnSiNi2A steel after 20 minutes of the Austenite process at 900°C, 230°C tempering for 60 minutes and then cooling to room temperature, is lathy Martensite + Bainite + retained Austenite (Figure 4). The measurement of magnetization indicates that the amount of retained Austenite can reach 15%. The retained Austenite is transformed in the process of annealing. After 200°C annealing, the fraction of Austenite is reduced to 12%; after 275°C annealing, it declines to 6~7% [8]. Figure 5 shows the structure of 30CrMnSiNi2A steel after tempering and 275°C annealing for 3 hours.

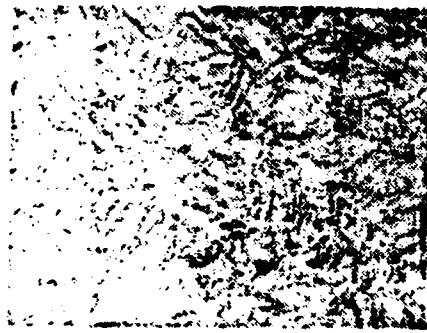
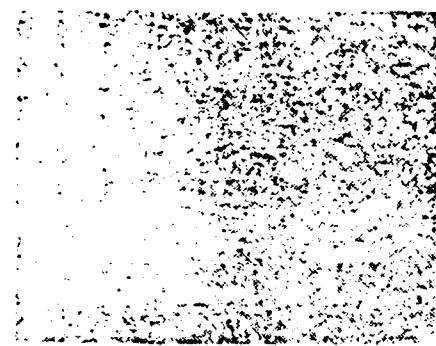


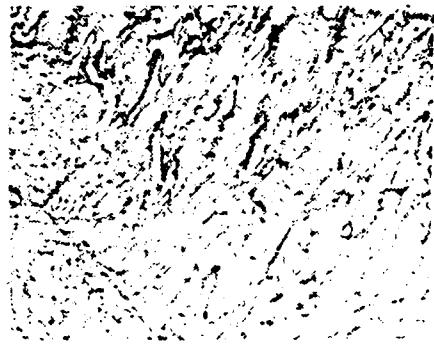
Fig. 3



Fig. 4



a)



b)

Fig. 5

Figure 3. The structure of the 30CrMnSiNi2A steel after 20 minutes of Austenite process at 900°C, oil-quenching, 200°C annealing for 3 hours 500X

Figure 4. The structure of the 30CrMnSiNi2A steel after 20 minutes of Austenite process at 900°C, 230°C tempering for 60 minutes, cooling to room temperature 9500 X (double complex type)

Figure 5. The structure of the 30CrMnSiNi2A steel after 20 minutes of Austenite process at 900°C, 230°C tempering for 60 minutes, cooling along with 275°C annealing for 3 hours

a) optical 500X      b) double complex type 9000 X

### Deformation Reliefs and Structure

After the appearance of cracks on samples through impact fatigue testing, stress concentration occurs near the tip of the crack, and tremendous plastic deformation occurs for materials in that region. This leads to so-called "deformation reliefs" on the sample surface, as shown in Figure 6a. As the crack propagates, it leaves behind deformation reliefs on both sides of the seam, as shown in Figure 7a.

As can be seen, in a high-strength Martensite steel, the plastic deformation near the tip of the crack is not like the one described in the macroscopic mechanics which considers a uniform distribution along a symmetrical surface (seam surface for this case). As a matter of fact, it is affected to a large extent by the structure of steel and the mechanism governing the plastic deformation.

The sample was polished and the "deformation reliefs" were removed. Metallic structure was exposed by etching the polished surface, as shown in Figures 6b and 7b. Comparing Figures 6a and 7a with Figures 6b and 7b, Martensite laths can be found right under the deformation reliefs. Both the orientation and the size of the deformation reliefs correspond exactly to those of the Martensite laths. On the basis of these observations, it is reasonable to suggest that the appearance of the deformation relief is due to the sliding of Martensite laths, which is an important mode of plastic deformation of the lathy Martensite structure.

Whether or not sliding exists between Martensite laths is closely related to the orientation of the Martensite lath bundles, and is related to the applied stress. When the Martensite lath bundle is nearly perpendicular to the crack, there will be no sliding between Martensite laths in the bundle (Figs. 6 and 7).

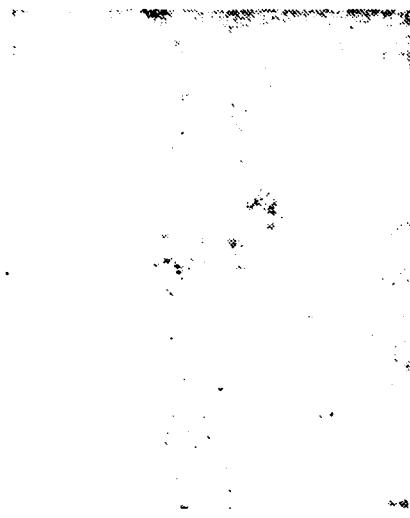


a)



b)

Fig. 6



a)



b)

Fig. 7

Figure 6. The deformation reliefs near the tip of a crack (a) and the structure (b). The sample undergoes 900°C Austenite process, oil-quenching and 200°C annealing. 500 X

Figure 7. The deformation reliefs on both sides of the seam (a) and the structure (b). The sample undergoes 900°C Austenite process, 230°C tempering 60 minutes and then cooling.

a) 500 X; b) 800 X

### Micro-Crack Initiation, Crack Propagation and Structure

When the amount of slip in the plastic zone around the tip of the crack exceeds a certain limit, a micro-crack is initiated along the boundary between Martensite laths (Figure 6). Under the action of alternating stress, the micro-crack in the plastic zone near the tip of the major crack is combined with the major crack, causing the propagation of major cracks (Figure 6). Accordingly, the major crack always tries to penetrate the boundaries between Martensite laths and propagates ahead. In the process of propagation, however, if the major crack faces a Martensite lath bundle perpendicular to the crack, it may also break the Martensite lath and continue to propagate (Figure 7b).

### Fatigue Crack Initiation Life

The fatigue crack initiation life is listed in Table 1.

Test results indicate that the 30CrMnSiNi2A steel, after oil-quenching and low-temperature annealing, has less retained Austenites, but has high yield stress and long FCIL. After tempering and low-temperature annealing, the steel will have more retained Austenites with a lower yield stress, thus shortening the FCIL. Apparently, more retained Austenites will lower the yield stress, and shorten the FCIL. This conclusion is identical to that reported in Ref. [6]. Again, it proves that the amount of the retained Austenites in a high strength Martensite steel has a large effect on the FCIL.

Table 1. The Fatigue Crack Initiation Life of 30CrMnSiN12A Steel.

Heat Treatment				Breaking ratio	No. 2	No. 2	Retained
System	kg/mm <sup>2</sup>	kg/mm <sup>2</sup>	%	work kg/mm/mm <sup>3</sup>	times	times	Austenite %
900°C, 20 minutes					5840	5371	~12
Austenite process,	173.0	105.0	46.5	112	5500		
230°C tempering					5340		
for 60 minutes,					5130		
cooling, 275°C					4940		
annealing for							
3 hours							
900°C, 20 minutes					6020	6145	6~7
Austenite process,	153.0	129.0	51.0	127	5750		
230°C tempering					6160		
for 60 minutes,					7200		
cooling 275°C					5970		
annealing for							
3 hours							
900°C, 20 minutes					4160	4160	/
Austenite process,	/	/	/	/	4200		
230°C tempering					3800		
for 60 minutes,					4560		
cooling, negative							
prestrain, 275°C							
annealing 3 hours							
900°C, 20 minutes	173.0	138.5	45.7	116	15100	11580	1~2
Austenite process					9200		
oil-quenched, 200°C					11000		
annealing 3 hours							
Heat treatment same					45000	/	/
as above + positive					>45000		
prestrain					>45000		

Footnote: Static strength and plastic indicator are from Refs. [16] and [17], data for breaking ratio work are derived from formulas in Ref. [2].

Table 1 shows that the local prestrain at the notch root has a great effect on the FCIL. Prestrain in the negative direction shortens the FCIL by over 30%, while prestrain in the positive direction prolongs the FCIL over 3 times. This is due to the residual pulling stress at notch roots, caused by the prestrain in the negative direction, thus shortening the FCIL. This is consistent with results reported in Ref. [7].

### DISCUSSIONS

After direct tempering or Martempering, the 30CrMnSiNi2A steel takes on a lathy Martensite structure. In addition, a certain amount of retained Austenites remains. Under isothermal tempering conditions, a small amount of Bainites remain, because Bainite transformation is likely to happen below the Mo point [9]. Following the order to tempering transformation [11], the 30CrMnSiNi2A steel, after Martempering and 200°C annealing, shows small carbides which are deposited in Martensite laths. In an isothermal tempering process, since the Martensite involves tempering once in an isothermal process, and a second martempering could occur during a slow cooling process, carbide deposition is observed in a structure which has experienced isothermal tempering and annealing (Figure 4). The deposition of carbides is even more evident with isothermal tempering as well as 275°C annealing (Figure 5). The retained Austenite may be sandwiched between Martensite (or Martensite and Bainite) laths like a thin film or lamella, as shown in Figures 4 and 5. Kong Mehkuang et al. have been using thin film samples to observe 18CrMn2MoBA steel structure under a transmission electron microscope [13]. They discovered a layer of carbon-rich structure sandwiched between Martensite (or Bainite) laths. The structure could be retained Austenite or an M-A structure. Thomas [12] studied the tempered structure of alloy steel containing

0.3% carbon with electro-optical methods. He confirmed a thin film of retained Austenite sandwiched between Martensite laths, only with the exception of individual steel categories (Fe-C-Mo).

Swarr and Krauss studied plastic deformation of lathy Martensite with little carbon [5]. They pointed out that the tempered lathy Martensite sample, after plastic deformation, grows a perfect dislocated cell structure in the lath, thereby confirming a crossing sliding generated in the Martensite lath. However, after annealing (400°C 1 minute) the sample with small carbide grains deposited in the Martensite lath does not have a dislocated cell structure in a Martensite lath after plastic deformation. The density of dislocations and its distribution have not significantly changed. Deformed crystals have not been observed either. Accordingly, we can consider that the plastic deformation of tempered Martensite lath is rather small; so is its contribution to plasticity.

As is well-known, the high-strength Martensite steel containing low carbon has a superior plasticity. According to Swarr and Krauss' work, we can conclude that there must be another plastic deformation model for the lathy Martensite structure. According to our observation on the "deformation reliefs" appearing in the plastic zone near the tip of the fatigue crack, we conclude that both the orientation and the size of the deformation reliefs correspond exactly with those of the Martensite laths. As a result, it is reasonable to propose another important plastic deformation model for a lathy Martensite structure, i.e., the relative sliding between Martensite laths. The contribution of relative sliding between Martensite laths to the plasticity is rather large. This is also a major factor for the high plasticity of the lathy Martensite structure.

This plastic deformation model also enables some experimental results to be explained. Since the resistance of plastic deformation of retained Austenite is low, the Martensite laths may slide relatively easily if the retained Austenite is sandwiched between Martensite (or Martensite and Bainite) laths. A low yield stress with the macroscopic mechanical performance occurs. Parker [14] has pointed out that a small plastic deformation may lead to a transformation of retained Austenite into Martensite. Additionally, this transformation may expedite the formation of fatigue cracks [15]. Under alternating stress, relative sliding of Martensite laths occurs near the notch root where stress concentrates. This will induce transformation of retained Austenite into Martensite and initiates the cracks along the boundaries between Martensite laths. Accordingly, the existence of retained Austenite initiates fatigue cracks and shortens the FCIL. The greater the amount of retained Austenite, the smaller the yield stress, and the easier it becomes for the crack to be initiated, and the shorter the FCIL.

It has been pointed out in Ref. [2] that the FCIL of steel grows nearly proportionally to its yield stress and breaking ratio work. However, present test results are unable to completely confirm the rule (Table 1). We think that this may be attributed to different maximum stress near the notch roots of the samples during two tests. Data in Ref. [6] were collected under the condition that the maximum stress near notch roots of the sample is far beyond its yield stress. The maximum stress near notch roots of the sample in our test is 155 kg/mm<sup>2</sup> which could be below the dynamic yield stress of steel for a oil-quenched and low-temperature annealed sample. On the contrary, for a sample which has been isothermally tempered, it might surpass its dynamic yield stress, or even surpass its elastic limit. By applying the foregoing viewpoint,

such a phenomenon can be reasonably explained. For an oil-quenched and low temperature annealed sample, a measurable plastic deformation region is not likely to be initiated near its notch roots in an impact fatigue test process. The plastic deformation, e.g., the relative sliding between Martensite laths, can be accumulated slowly only with increasing stress cycles, so the crack is not likely to appear and the FCIL is prolonged. For a few samples which have been isothermally tempered (230°C, 60 minutes), especially for the one with the isothermal tempering and 200°C annealing treatment, a measurable plastic region appears near the notch roots (following the Irwin calculation method). As a consequence, the Martensite laths in the plastic region will slide relatively and lead to the appearance of cracks, and shorten the FCIL. It should be remembered that the relation between the FCIL and breaking ratio work is not clear at this point. The proportional relation between them may be conditional.

The mechanism which reflects the effect of local prestrain at notch roots on the FCIL is still unclear. It bears further investigation.

In summary, investigations on the structure of Martensite laths and enhancement of its combination strength are important in order to upgrade its yield stress and prolong its FCIL.

#### CONCLUSIONS

1. The structure of 30CrMnSiNi2A steel is mainly lathy Martensite temperature annealing after tempering or following isothermal Martempering. A certain amount of retained Austenite remains.

2. The "deformation reliefs" appear in the plastic zone around the tip of the fatigue crack. Both the orientation and the size of the deformation reliefs correspond exactly with those of the Martensite laths.

3. It is reasonable to suggest that the appearance of the deformation relief is due to the sliding of Martensite laths, which is an important mode of plastic deformation of the lathy Martensite structure. It is also one of the major reasons why lathy Martensite structures have high plasticity.

4. The greater the amount of retained Austenite, the smaller the yield stress, and the shorter the FCIL. This may be attributed to the distribution of retained Austenite on the boundaries of Martensite laths, which affects the structure on the boundaries and combination strength.

5. The local prestrain at notch roots has a large effect on the FCIL. Prestrain in the positive direction prolongs the FCIL, while prestrain in the negative direction shortens it.

#### REFERENCES

- [1] Barnby, J. T., Dinsdale, K. D., and H. Initiation, Conference on the Mechanics and Ph. Cambridge. Paper 26.
- (2) Zheng Ziulin, Concerning service life without fatigue cracking, unpublished.
- ... Lankford, J., and Kusenberger, F. N., Initiat 4340 Steel. Met. Trans., Vol. 4, No. 2, Feb. 1973, pp. 553~559.

[4] Grosskreutz, J., Strengthening and Fracture in Fatigue (Approaches for Achieving High Fatigue Strength), Met. Trans., Vol. 3, No. 5, May 1972, pp. 1255-1262

[5] Swarr, Thomas., and Krauss, G., The Effect of Structure on the Deformation of As-Quenched and Tempered Martensite in an Fe-0.2 pct Alloy, Met. Trans., Vol. 7, No. 1, Jan. 1976, pp. 41-48.

[6] Northwestern University of Technology, Teaching and Research Sections of metallic materials and heat treatment processes on overall mechanical properties of 30CrMnSiNi2A steel, unpublished.

[7] Radhakrishnan, V. M., and Baburamani, P. S., Initiation and Propagation of Fatigue Crack in Pre-strained Material, Inter, J. Fracture, Vol. 12, No. 3, June 1976, pp. 369-380.

[8] Northwestern University of Technology, Teaching and research sections of metallic materials and heat treatment. Research on measuring the remaining austenite and its tempering changes upon zonal isothermal quenching of 30CrMnSiN;2MoA steel martensity, unpublished.

[9] Northwestern University of Technology, Teaching and research sections of metallic materials and heat treatment. Analysis of structural changes during isothermal quenching and tempering of 30CrMnSiN;2A steel, unpublished.

[10] Zheng Ziulin, Research related to the use of an impact fatigue tester to measure fatigue life without cracking of steel. Materials to be published soon.

[11] Speich, F. R., and Leslie, W. C., Tempering of Steel, Met. Trans., Vol. 3, No. 5, May 1972, pp. 1043-1054.

[12] Thomas, Gareth., Retained Austenite and Tempered Martensite Embrittlement, Met. Trans., Vol 9A, No. 3, Mar. 1978, pp. 439-450.

[13] Kang Mokuang, Guan Dunhui, Guo Chengda, et al. The M-A structure in low-alloy, high-strength structural steel. Materials soon to be published.

[14] Parker, Earl. R., Interrelation of Compositions, Transformations Kinetic, Morphology, and Mechanical Properties of Alloy Steel, Met. Trans., Vol. 3A, No. 7, July 1977, pp. 1025-1042.

[15] Baudry, G., and Pineau, A., Influence of Strain-induced Martensitic Transformation on the Low-cycle Fatigue Behavior of a Stainless Steel, Mater. Sci. Engr., VOL. 28, No. 2, May 1977, pp. 229-242.

[16] Hu Guangli, Research on heat treatment processes of 30CrMnSiNi2A high-strength structural steel (Part 1)- Structure and properties after oil quenching and tempering, unpublished.

[17] Northwestern University of Technology, Teaching and research sections of metallic materials and heat treatment, unpublished.

THE COMPUTATION OF INTEGRAL-TYPE FLEXURE HINGE ASSEMBLY  
OF DYNAMICALLY TUNED GYROSCOPES

Mei Shuo-ji

Translation of "Song Li Tiao Xie Shi Two Luo Nao Xing Zhi  
Cheng (Zheng Ti Shi) De She Ji Ji Suan," from Xi Bei Gong  
Ye Da Xue Lun Wen Xuan (Part One), 1979

## Summary

# The Computation of Integral-Type Flexure Hinge Assembly of Dynamically Tuned Gyroscopes

*Mei Shuoji*

The calculation of flexure hinge assembly is an important problem in the design of a dynamically tuned gyroscope. In order to satisfy the tuning requirements of the gyroscope, the calculation of the angular stiffness of the assembly must be highly accurate. In order to eliminate the  $g^2$  drift error, the axial stiffness and the radial stiffness of the assembly must be equal. The purpose of this paper is to investigate the calculation methods for finding the various stiffnesses of integral-type flexure hinge assembly.

The hinge assembly has a complex shape and consists of two coaxial parts: an inner hinge unit and an outer hinge unit. Each hinge unit has four neck-shaped elements working as flexure bars. For the convenience of calculation, three assumptions are made as follows: (1) the deflection of the hinge assembly comes only from the deformation of the thin neck-shaped elements, the other parts of the assembly being regarded as rigid bodies; (2) the stiffness of the assembly depends only on the manner of combination of the eight single elements of the two units, such as in series or in parallel; (3) only the axial-stress effect and the bending effect are considered, the shear effect being neglected for being much smaller than the above effects.

The above assumptions reduce the calculation work to the finding of the stiffnesses along different axes of a single element under various loading conditions. In this way, the effect of structural parameters can be easily analyzed and parameter values can be selected to obtain optimum performance.

By employing above method, accurate results are obtained for angular stiffness calculation, the error of computed results being within 1% as compared with test data. On the other hand, although tests confirm that axial and radial stiffnesses are quite close, the error of the computed results is too large as compared with test data, thus indicating that the assumptions made are not appropriate in this case. The results, however, can still be helpful suggestions in the selection of proper parameter values in the design of flexure hinge assembly whose axial and radial stiffnesses should equal.

THE COMPUTATION OF INTEGRAL-TYPE FLEXURE HINGE ASSEMBLY  
OF DYNAMICALLY TUNED GYROSCOPES

Mei Shuo-ji

SUMMARY

/133

The calculations which are related to the design of flexible hinge assemblies are one of the important problems involved in the design of dynamically tuned gyroscopes. In order to satisfy the requirements of gyroscopic tuning, the degree of the angular displacement stiffness which is designed into the piece of equipment must have a high degree of accuracy. In order to eliminate the error associated with  $g^2$  drift, the axial stiffness and radial stiffness of the hinge assembly of the gyroscopes being considered should be equal. Besides this, there are naturally quite a few other requirements which need to be considered. This article only addresses itself to the discussion of the actual structure of one integral type hinge assembly as well as to an investigation of the methods for designing the angular displacement stiffness, the axial stiffness and the radial stiffness specifications for the gyroscopes being discussed.

In this section, we will consider the following:

(1) Tuning Requirements

The conditions which govern the tuning of dynamically tuned gyroscopes are shown in the equation set out below (1), that is,

$$K_x + K_y = (a + b - c)N^2 \quad (1)$$

In this equation,  $K_x$  and  $K_y$  are the angular displacement stiffness for the x and y axes of a hinge assembly. a, b, and c are the inertias of equatorial rotation and mechanical inertia

for the oscillatory equilibrium ring of flexible hinge assemblies.  $N_0$  is the operational speed of rotation for the gyroscope, that is to say, the tuned speed of rotation.

Obviously, if one is talking about a gyroscope for which the mechanism and dimensions have already been determined, then, the angular displacement stiffness,  $K_x$ , and  $K_y$ , as well as the rotational inertias, a, b, and c of the oscillatory equilibrium ring, and other numerical values of a similar kind are already known as well. In this sort of situation, it is possible to make use of a method for changing the operational speed of rotation of the gyroscope involved in order to obtain the specific speed of rotation which is required to satisfy equation (1). At this speed of rotation, when the gyroscope rotor has an angle of rotational offset in relation to the body of the gyroscope, the oscillatory equilibrium ring of the assembly forms a moment of dynamic force, and the value of this moment of force is equal to the moment of elastic force which is produced by the hinge assembly rods. The directions of these two moments of force are the opposite of each other, and so they are in balance with respect to each other. This causes the gyroscope to enter a state in which its moments of force cancel each other out and one may say that the gyroscope is tuned. After one reaches this point, on the basis of the conditions which are given for the designing of the gyroscope, all the operational speeds of rotation are either given or selected. In this way, once the designing has begun, then, it is necessary to be concerned and to take into consideration the adjustment values required. After this is done, it is still necessary to obtain the angular displacement stiffness,  $K_x$ , and  $K_y$ , as well as the numerical values for the oscillatory equilibrium ring rotational inertias, a, b, and c, which are appropriately determined for the speeds of rotation selected. It is also necessary to take the relationships between these quantities and satisfy the fixed adjustment conditions of equation (1) to

the degree of accuracy required. From this it can be seen that the degree of accuracy in the calculation of the angular displacement stiffness is extremely important to the design of the gyroscope being considered. This is the problem which demands first attention during the design calculations for hinge assemblies.

(2) The lowering of the requirements on  $g^2$  drift error

In the case of inertial gyroscopes, the problem of lowering the drift error associated with  $g^2$  is one which also requires attention. That is to say that, when the gyro rotor is subjected to axial and radial accelerations, it is necessary to satisfy the requirements for rigidity in equal amounts in all directions in order to facilitate eliminating the  $g^2$  drift error which is caused by the acceleration of the aircraft or carrier. Therefore, what is necessary is to carry out calculations of the axial and radial stiffnesses of hinge assemblies, and, from these, find a way to select numerical stiffness values, and so obtain structures with equal stiffness values in order to lower the  $g^2$  drift error. This is another question which must be considered when dealing with the design calculations for hinge assemblies.

Besides the two problems which were raised above, among the other requirements which must be considered is the question of the energy to power the activation electrodes of the gyroscope in question as it is transferred from the hinge assembly. Load bearing strength and fatigue strength as well as the ability to resist the absorption of energy from blast and oscillation as it exists in the environment of the gyroscope all contribute to the maintenance of the normal operation of the gyroscope. Besides this, it is also necessary to give consideration to the capabilities of facilities available for the manufacture of gyroscopes as well as the design planning for

the structure of the assemblies being considered. However, these questions are not discussed in this article.

## II. INTEGRAL TYPE FLEXURE HINGE ASSEMBLIES

At present, there are two different structural types in use in dynamically tuned gyroscope flexure hinge assemblies. These are the composite type and the integral type. This article only discusses the question of the design calculations for integral type hinge assemblies.

What is meant by an integral hinge assembly is a three-ring bearing component which has undergone final machining, is made out of metal, and is divided into three sub-components which are connected by hinge rods of low elasticity and each of which has some freedom of movement at right angles. After each of the three mounting rings is mated to the activating electric motor of the gyroscope and to the gyroscope's rotor, the rotor is fitted onto the base (casing), and one then has the normal configuration of the gyroscope, which allows for freedom of movement. Obviously, this sort of orthogonal freedom of movement comes from the action of the hinge rods; however, because of the need for satisfying precise requirements of rotor axis rotation and low elasticity, the geometrical configuration of these rods and their dimensions place limits on other load bearing capabilities, and this causes the assembly to be almost incapable of sustaining loading from any direction other than that of the centerline of the hinge rods. Because of this fact, most of the practical bearing assemblies are designed as two components; moreover, this causes the hinge rods involved to have different placements. The hinges of one of the components mainly bear the loading along the axis of the gyroscope. The other component's hinge rods mainly bear the radial loading of the gyroscope involved. When the two sets of component rods are formed together into one integral

body, then, they become one integral hinge assembly capable of taking loading from various directions. However, speaking of the matter of the angle of rotational deviation which is involved in the fitting of the gyroscope rotor to the main casing, the operation of the two components of this type of integral hinge assembly is exactly the same as that of a single component.

Figure 1 is a structural diagram [2] of one type of integral hinge assembly. From the figure it can be clearly seen that there is a definite relationship between the hinge rods of each assembly component and the various ring supports of the assembly as a whole. Obviously, the hinge rods of the interior and exterior components are all separated after circles of equal diameter are bored in the walls of the component to form a narrow neck of a definite configuration and are fitted with passageway apertures configured to a different pattern. The main reason for this is this. These hinge rods can also be called narrow neck components. On the basis of the straight axial arrangement seen in the illustration, one can see that the radial loading along the x and y axes of the gyroscope is supported by the interior components. (two characters unreadable) the axial loading along the z axis is, then, born by the exterior component. After the interior and exterior components are connected together, they (two characters unreadable) they are not only capable of bearing loading along different axes, they are capable of maintaining (two characters unreadable) the freedom of movement required in the fitting of the gyroscope rotor to the base's x and y axes.

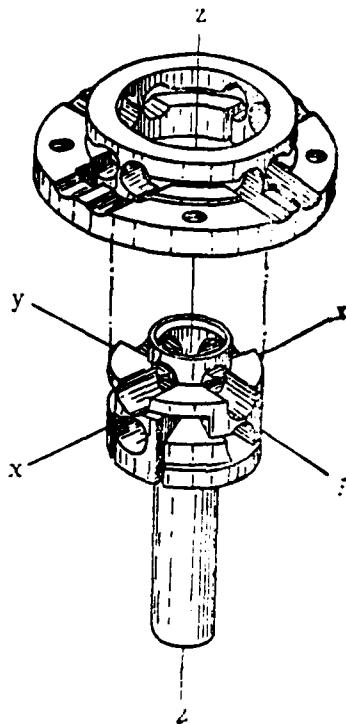


Figure 1. An Exploded Diagram of the Interior and Exterior Components of Integral Flexure Hinge Assembly

### III. DESIGN CALCULATIONS FOR AN INTEGRAL TYPE FLEXURE HINGE ASSEMBLY

This article only deals with calculations for angular displacement stiffness, axial stiffness and radial stiffness as these quantities relate to the calculations involved with integral type flexure hinge assemblies. Other calculation problems are not considered here.

#### (1) The Calculation of Angular Displacement Stiffness

From Figure 1 we already know that the angular displacement stiffness of an integral type assembly around its x and y axes is composed of the angular displacement stiffnesses of

the two coaxial parts of the two components in the narrow neck. The assumptions involved in this are as follows.

(1) Under the effect of loading (bending moment), only the narrow neck parts within the range of the diameter of the holes give rise to deformations. The other parts of the hinge assemblies can be taken to be rigid bodies.

(2) The deformation rigidity of integral assemblies is only determined by taking together the different modes experienced by the eight narrow neck components of the two structures in the hinge assembly (for example, in series mode deformations and parallel mode deformations).

For the present, let  $K'$  interior angle represent the angular displacement rigidity of the flexure axis around a single narrow neck component of the interior component of the flexure hinge assembly. Let  $K'$  exterior angle represent the angular displacement rigidity of a single narrow neck component of the exterior component of the flexure hinge assembly. Of course, in this case, we take the dimensions of each of the narrow neck components of a structure of the hinge assembly to be the same. After we do this, the angular displacement rigidity around the x or y axis of an integral hinge assembly, i.e.  $K_x$  angle or  $K_y$  angle, can then be calculated from a combination of these quantities, that is,

$$K_x \text{ angle} = K_y \text{ angle} = 2(K' \text{ interior angle and } K' \text{ exterior angle}) \quad (2)$$

The reason for this is the fact that the dimensions of the operating elements of the narrow neck components are very small -- only a few hundredths of a millimeter -- and the dimensions of other elements are all much larger than that.

Because this is the case, after one ignores the angular displacement deformations of all the components besides the narrow neck components, the error which arises in the calculations is not large enough to be evident.

Concerning the angular displacement rigidity of the narrow neck components,  $K'$  interior angle or  $K'$  exterior angle, it is possible, on the basis of the effect of the pure bending moment of the flexure curve around the axis, to use the relationship between angular deformations of changes in rod cross sections within the range of the radii of the holes in order to make calculations. In Appendix I, one may find given a method for computing the angular displacement rigidity of the narrow neck components in this type of situation.

## (2) The Calculation of Axial Rigidity

It has already been pointed out before that, if one is talking about the type of integral flexure hinge assembly which is shown in Figure 1, then, the loading on it in the direction of the  $z$  axis is supported by the four narrow neck components running right along the direction of the  $z$  axis which belong to the exterior subassembly of the hinge assembly.

Obviously, in a situation in which loading along the  $z$  axis is born by the exterior sub-assembly, it is possible to use the relationships shown in Figure 2 in order to explain the situation. In this illustration, we have made use of extended (or compressed) springs in order to represent narrow neck components under the effects of loading directly along their axes; moreover, it draws, in the same plane, four narrow neck components distributed in a radial direction. This sort of arrangement will not influence the analysis and results concerning the problems being discussed.

From Figure 2, it can be seen that, when ring III of the assembly is under the influence of a load,  $P$ , (this ring is in direct contact with the rotor of the gyroscope), then, due to the fact that the narrow neck components A, A and B, B both form a symmetrical distribution along the  $z$  axis, the two narrow neck components, A-A, between the upper ring, ring III, and the oscillation equilibrium ring, ring II, respectively experience the effects of the load  $P/2$  directly along their axis lines, and are put into an extended configuration. At the same time, by means of a transmission through the oscillation equilibrium ring, ring II, the two narrow neck components, B-B, between the oscillation equilibrium ring, ring II, and the bottom hinge assembly ring, ring I, (this ring is in contact with the axis of the electric activation motor), only respectively feel the effects of the load  $P/2$  directly along their axis lines, and this puts these rings into a compressed configuration. In this way, the four narrow neck components are first arranged in the parallel pairs, A-A and B-B. After this, AA-BB becomes the arrangement for series operation. The reason for this is that the geometrical dimensions of all the narrow neck components of similar assemblies are designed to be similar. Therefore, when the axial rigidity of a single narrow neck component directly along the line of its axis is represented by the use of  $K'_{\text{outside axis}}$ , then the axial rigidity of the exterior component,  $K_{\text{outside axis}}$ , can be solved for by using the formula given below, that is,

$$\frac{1}{K_{\text{outside axis}}} = \frac{1}{2K'_{\text{outside axis}}} + \frac{1}{2K'_{\text{outside axis}}} \quad (3)$$

or,

$$K_{\text{outside axis}} = K'_{\text{outside axis}} \quad (4)$$

The results from Equation (4) show that, when one considers the axial rigidity of exterior components one by one, they are simply equivalent to the axial rigidity,  $K'$  outside axis of a narrow neck component. Concerning the axial rigidity,  $K'$  outside axis of a single narrow neck component, it is possible to calculate, on the basis of the tensile loading which is applied, the tensile deformation of a rod cross section under tension within the range of the radius of the penetration aperture. Appendix II presents a method of calculation for obtaining the rigidity directly along the line of the axis of a narrow neck component in this type of configuration.

From the external components which are shown in Fig. 1, one can see that, when the effects of a load  $P$  are felt along the z-axis of support ring III (close to the gyroscope rotor), this load  $P$  will first cause force to be applied to the two narrow neck components, C-C, between upper ring III and the oscillation equilibrium ring II. After this, this same force will be transmitted through oscillation equilibrium ring II, and it will cause the two narrow neck components between the oscillation equilibrium ring II and the lower support ring I (quite close to the axis of the activation motor) - that is, D-D - to have force exerted on them. Due to this, the two sets of narrow neck components can also be taken to form two sets of two parallel components each, or can be considered in another way as operating as a set of series components. The letters for these two situations are C-C and D-D in the first case, and CC-DD in the second case. Fig. 3 illustrates the loading situation when loading is applied axially along the internal component supports. To formulate a graphic representation of this phenomenon, we use spring plates to represent the various narrow neck components. These components are also shown in a plane illustration.

It can be seen from this illustration that, when the loading affecting the length of the z-axis is distributed to the various narrow neck components, it is possible to establish that the lines along which this loading appears all pass through the intersection of the axes of the various narrow neck components.

Similarly to the calculation method for the angular displacement rigidity factor, one must also consider the fact that a support assembly operates as a combination of its interior and exterior components. Because of this, it is not only the narrow neck component of the exterior sub-assembly which bears axial loading along its z-axis in a direct line with that axis. The narrow neck component of the interior sub-assembly also supports loading directly along the x-axis line. Without any question, when one is figuring the axial rigidity of a hinge assembly as a whole, one ought also to take the axial rigidity of the internal and external sub-assemblies into account because it is only in this way that one can make a reasonable approximation to the actual situation.

From the interior and exterior sub-assemblies which are shown in Fig. 1, it can be seen that, when the loading which acts along the x-axis is distributed among the various narrow neck components, it is possible to recognize the fact that the lines along which the loading acts all pass through the intersection of the axes of the various narrow neck components. The actual situation is this. The loading is transferred through the supporting ring assemblies and then transmitted by one end

of the narrow neck components. Because of this situation, speaking from the point of view of the narrow neck components, it is possible to take them and see them as cantilever beams (rods). When they have exerted on them the effects of a bending moment formed by a load,  $P'$ , for which the distance from the pivot point is  $2R$ , these components also have exerted on them the effects of a moment of couple the direction of which

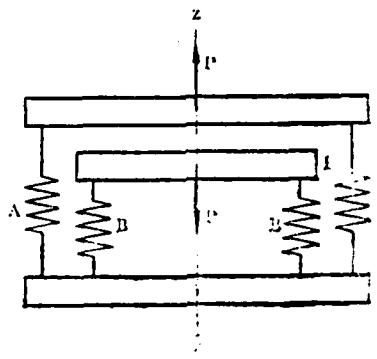


Figure 2. External Structural Form of Model  
the Effects of Axial Loading.

is opposite to that of the bending moment which was discussed above and has a numerical value of  $PR$  (Figure 3b). The deformation rigidity at the intersection of the axes, which is created from the bending moment  $2P'R$  and the moment of couple  $PR$  makes it possible, on the basis of the loads given in Appendix III, to calculate the deformational rigidity from the rigidity at the intersection of the axes of the narrow neck components involved. However, it must be pointed out that what we are concerned with is the axial rigidity of the interior components, that is to say, the deformation rigidity along the  $x$  axis of the support assembly, and calculations should be made with this idea in mind.

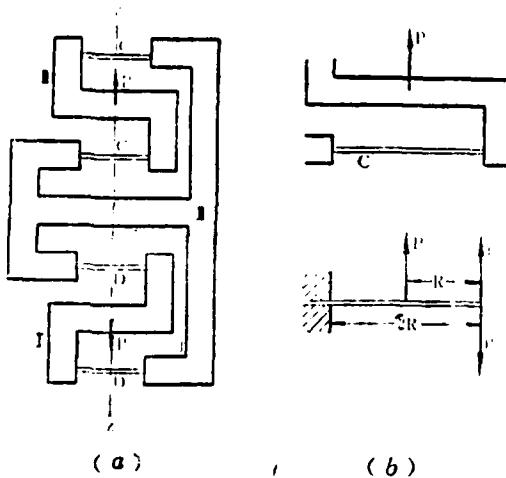


Figure 3. An Interior Component Under the Effects of Axial Loading.

After one obtains the rigidity of the intersection of the axes,  $K'_{\text{inside axis}}$ , for each of the single narrow neck components of the interior sub-assembly, then, in the same way, it is possible, on the basis of the series and parallel relationships of the various narrow neck components of a sub-assembly to solve for the axial rigidity of the interior component,  $K_{\text{inside axis}}$ , as follows, that is,

$$\frac{1}{K_{\text{inside axis}}} = \frac{1}{2K'_{\text{inside axis}}} + \frac{1}{2K'_{\text{inside axis}}} = \frac{1}{K'_{\text{inside axis}}} \quad (5)$$

or, to put it another way,

$$K_{\text{inside axis}} = K'_{\text{inside axis}} \quad (6)$$

Equation (6) shows that, similar to the axial rigidity of the exterior subassemblies, when one is giving consideration to the various separate axial rigidities in an interior sub-assembly, these are also only equivalent to the individual

axial rigidities,  $K'$  inside axis, of the narrow neck components within the assemblies being considered. On the basis of the actual situation as far as exerted forces are concerned, each of the narrow neck components of the interior sub-assembly also has exerted against them shear loading of a definite magnitude, and this loading can also be solved through the calculation process. The only thing is that the degree of this shear loading, when compared to the magnitude of other forces involved, is very small in numerical terms, and can be ignored. Due to this fact, the axial rigidity of the hinge assembly as a whole,  $K_{\text{axial}}$ , can be solved for by using the equation below, that is,

$$K_{\text{axial}} = K_{\text{inside axial}} + K_{\text{outside axial}} \quad (7) \quad /137$$

### (3) The Calculation of Radial Rigidity

It has already been pointed out that, when one is speaking from the point of view of the integral-type hinge assemblies we have been discussing, the radial rigidities of this type of assembly are mainly supported by the narrow neck components of the interior and exterior sub-assemblies. Figure 4 is a drawing of the nature of the placements of the various narrow neck components in an equatorial plane, when the interior sub-assembly is being subjected to the effects of radial loading. It can be seen from this illustration that the axis of bending of the narrow neck component, C, C, is identical to the y axis, and that the axis of bending of the narrow neck component, D, D, is identical to the x axis. Now, let us assume that the line of the effects from the radial loading, P, of the gyroscope rotor lies along the line of support in the direction of the y axis. Obviously, the nature of the loading will be different from that for different narrow neck components such as C and D. If we talk about the narrow neck component C, for a moment, although the actual loading is transferred through one end of a narrow neck component, it is possible to consider that the line of effect for

loading passes through the axis of bending of the component involved, and calculations similar to the ones used to figure the axial rigidities of the interior sub-components and discussed in (2) above, can also be applied as calculation methods for the corresponding radial rigidities,  $K'_{\text{radial interior}}$ . All that needs to be added to the calculations is the fact that the direction of the loading on the narrow neck components involved are not the same. In this case, the effects of the loading travel along the axis of bending of the components involved. In Appendix IV, one finds presented a method of calculation for the radial rigidity,  $K'_{\text{radial interior}}$ , in the form of its effects along the bending axis of the narrow neck components involved. In the case of the narrow neck component, D, under the influence from effects of loading in the direction of the y axis, there will be caused a deformation right along the direction of the axis involved, and the corresponding radial rigidities,  $K''_{\text{radial interior}}$ , can be added to the calculation methods presented in Appendix I.

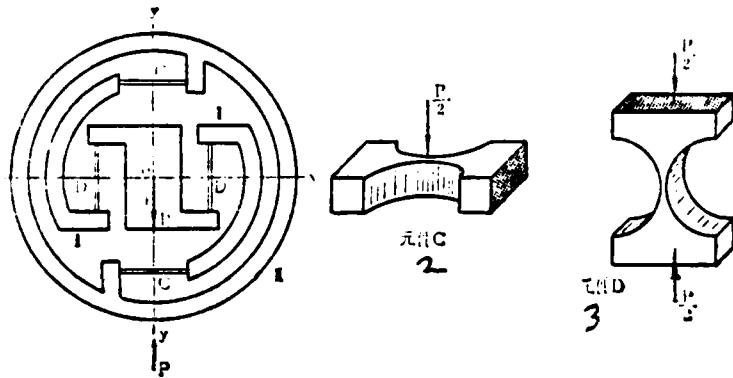


Figure 4. The Calculation of the Radial Rigidity of the Interior Sub-assembly. 2. Component C. 3. Component D.

Under the effects of radial loading, the combined arrangement of the various narrow neck components of the interior sub-

assembly can be found in the form of C-C, D-D arranged in parallel two at a time or in the form of CC-DD arranged in series. If one is only talking about the narrow neck components C and D, then, the directions of loading involved are not the same. Naturally, the rigidities of the two cases,  $K'$  radial interior and  $K''$  radial interior, will also produce calculation results which are not the same either. Depending on what the relationship is between the two types of narrow neck components, it can be possible to make use of the formula which follows in order to calculate the solution for  $K_{\text{radial interior}}$ . This formula is

$$K_{\text{radial interior}} = \frac{2K'_{\text{radial interior}}K''_{\text{radial interior}}}{K'_{\text{radial interior}} + K''_{\text{radial interior}}} \quad (8)$$

In situations similar to the ones discussed in (2) above, one should also include in the considerations the radial loading energy which is supported by the exterior sub-assembly concerned. Because of this, in Figure 5 one finds presented the placement relationships for the various narrow neck components of the exterior sub-assemblies involved when under the effects of radial loading along the y axis. Obviously, the two pairs of narrow neck components, A-A and B-B, can be operated both in the two-by-two parallel arrangement and in the AA-BB series arrangement. If one takes  $K'_{\text{radial exterior}}$  to represent the radial rigidity  $K'$  on the narrow neck component A, then, the radial rigidity of the exterior sub-assembly,  $K_{\text{radial exterior}}$ , can be found by using the equation below, that is,

$$K_{\text{radial exterior}} = \frac{2K'_{\text{radial exterior}}K''_{\text{radial exterior}}}{K'_{\text{radial exterior}} + K''_{\text{radial exterior}}} \quad (9)$$

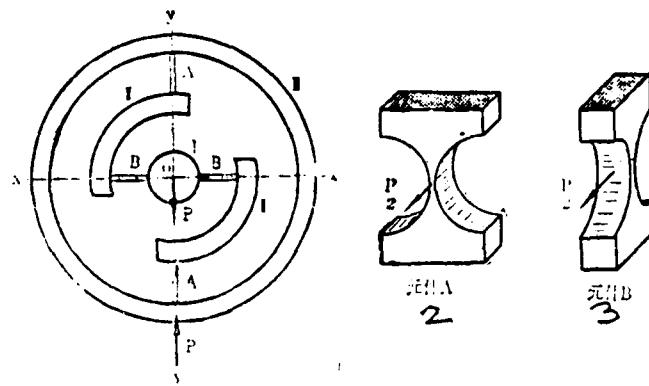


Figure 5. The Calculation of the Radial Rigidity of the Exterior Sub-assembly. 2. Component A, 3. Component B.

On the basis of the situation illustrated by Figure 5, it can be seen that the radial rigidity of the narrow neck component A of the exterior sub-assembly,  $K'$ radial exterior, is determined by the deformation which is caused by the loading along the axis of bending. The calculation method for this case can be seen in Appendix IV. If we are speaking of the radial rigidity of the narrow neck component B, that is,  $K''$ radial exterior, then, it is known that this rigidity is determined by the deformation caused by the loading along the cross axis. The method of calculation for this is found in Appendix III.

In the cases which we have just considered, we have always ignored the shear rigidity of the narrow neck components involved, for the reasons which were presented above. Due to this fact, the radial rigidity of the hinge assembly as a whole,  $K$ radial, can be calculated by using the equation presented below, that is,

$$K_{\text{radial}} = K_{\text{radial interior}} + K_{\text{radial exterior}} \quad (10)$$

#### IV. A CRITIQUE OF THE RESULTS OF CALCULATIONS

Concerning a hinge assembly of the structural type discussed above, after one makes a comparison between the data obtained on the structure from calculations and the measurements taken during experimentation, one discovers that:

(1) If one is speaking about the angular displacement rigidity of a structure, then, the numerical values which were calculated for the assembly as a whole were quite close to the results obtained through experimental measurements. This demonstrates that, on the basis of the assumed conditions presented above, the calculated values are reliable and relatively precise. If one takes the case in which an integral type of hinge assembly is made out of alloy materials of a constant elasticity, then, the radius of the eight pairs of drilled apertures in the interior and exterior sub-assemblies is  $D=2.5$  mm. The thickness of the narrow neck components at the thinnest place in the aperture involved is  $\Delta=0.05$  mm. The width of the interior sub-assembly,  $b_{\text{interior}}=1.75$  mm. The width of the exterior sub-assembly,  $b_{\text{exterior}}=1$  mm. The calculated angular displacement rigidity of the assembly,  $K_{\text{angular(calculated)}}=0.105$  g.cm/degree of angle, and the experimental measurement for the same angular displacement rigidity of the assembly,  $K_{\text{angular(experimental)}}=0.106$  g.cm/degree of angle. This represents an error of less than 1%. This type of degree of accuracy was achieved under conditions of precise rotational speed, which caused the gyroscope to smoothly enter into a state of tuned equilibrium, and offer an excellent foundation for the measurements. However, it should be pointed out that, the data presented above was obtained in tests on a small experimental sample, and we still lack a sufficiently large body of experimental data to offer more significant statistical data.

Of course, even if the analysis and calculations have relatively high degrees of accuracy, it still does not mean that, in the design of the structures for the hinges assemblies we have been discussing, it is possible to ignore the adjustment components needed in order to adjust these assemblies. In fact, due to various types of reasons, there can exist considerable overall differences between the calculated data and the experimental data, and it is necessary to very carefully satisfy the adjustment conditions given for equation (1), and then, it is necessary to make small necessary adjustments in the initial design considerations relating to the rotational inertia of the oscillatory equilibrium ring of the assembly involved.

(2) After doing a comparison between the calculated numerical values for the axial rigidity and the radial rigidity and the experimentally measured values for the same quantities, it can be seen that, although actual axial rigidities and radial rigidities are relatively close, they are still much smaller than the calculated numerical values. This indicates clearly that there are still problems with taking the methods presented earlier and using them to figure numerical values for these two quantities. This procedure still needs to wait on future research. There are two possible reasons for this. One reason is that, in this type of situation, consideration is only given to the deformations of narrow neck components within the scope of the drilled aperture radii, and these calculations are not extended to take in deformations in other components. The reason for this is that the dimensions of the hinge assemblies as a whole are relatively small and, speaking from the point of view of the axial rigidities and radial rigidities of these assemblies, the assumption that the narrow neck components of the hinge assemblies can be viewed as being rigid bodies is not appropriate, and this assumption necessarily

brings with it relatively large errors. The calculation of the quantities we just discussed is different from the case of the angular displacement rigidity calculations. In the case of these calculations, the angular displacement rigidity of the axis of bending is basically very small, and when a comparison is carried out, and we ignore the deformations in other components, its influence is not large. Another reason is the problem of experimental measurement. The theory and equipment used in experimental measurements of axial and radial rigidities still need to wait on further discussion and improvement.

However, the related calculations of axial and radial rigidities are certainly not without significance. What must be paid attention to in the design of these is the requirement to make the axial rigidity and the radial rigidity equal and to reach a state in which the rigidity of the various directions is equal. This is important in order to eliminate  $\xi^2$  drift error. This requirement, on the basis of the experimentally measured results, is still satisfied only to a certain extent. Because of this, it is possible in the calculations mentioned above, to analyse the influences of the related parameters, and to change them in order to make it possible to make the hinge assemblies which are designed even more adequately in compliance with the design requirement for equal rigidity in all the various directions.

#### APPENDICES

##### I. THE CALCULATION OF THE ANGULAR DISPLACEMENT RIGIDITY OF NARROW NECK COMPONENTS UNDER THE INFLUENCE OF THE PURE BENDING MOMENT OF THE AXIS OF BENDING

The narrow neck component which is shown in Figure I is formed from the walls of hinge assemblies with a pair of holes drilled in them with equal radii of  $2R$  and at a thickness of  $b$ .

The thickness of the thinnest section of this narrow neck component (which is the point located by the joining of the centerlines of this pair of drilled apertures) is  $\Delta$ . The calculation width is the normal value, and equal to  $b$ . In the illustration, the d-d axis is the vertical axis of this narrow neck component. The q-q axis is its cross axis, and the f-f axis is its axis of bending.

Under the influence of a pure bending moment of the bending axis, the narrow neck component will bend its axis of bending, f-f, and produce an angular deformation. This is the angular displacement which needs to be calculated. The numerical value of this angular displacement depends on the size of the moment of bending, the geometrical dimensions of the narrow neck component and the characteristics of the materials used. Under the influence of a bending moment  $M$ , let us select a differentiation length on the surface of the component,  $dx$ . If we do so, then, it is possible to calculate the corresponding differential angle of deformation,  $d\theta$ , according to this equation.

$$d\theta = \frac{M}{EJ_z(x)} dx \quad (I-1) /139$$

In this equation,  $M$  is the bending moment affecting the axis of bending.  $E$  is the modulus of elasticity of the material involved, and  $J_z(x)$  is the moment of inertia of the cross section of the component perpendicular to the  $z$  axis centerline of the plane of the moment of bending. In a case when the cross section of deformation is along the  $x$  axis, then,  $J_z(x)$  will vary with changes in  $x$ , that is to say,  $J_z(x) = bh_1^3(x)/12$ . In this case,  $b$  is the width of the component, a constant value, and  $h_1(x)$  is the height of the component. If we let the half cross section height of the component be  $h(x)$ , then, we have  $h_1(x) = 2h(x)$ . On the basis of this, we can get the relationship  $J_z(x) = 2bh^3(x)/3$ .

According to the coordinates used in Figure I, the half-height of the cross section,  $h(x)$  can be written in the following form, that is,

$$h(x) = (R + \Delta/2) - \sqrt{2Rx - x^2} \quad (I-2)$$

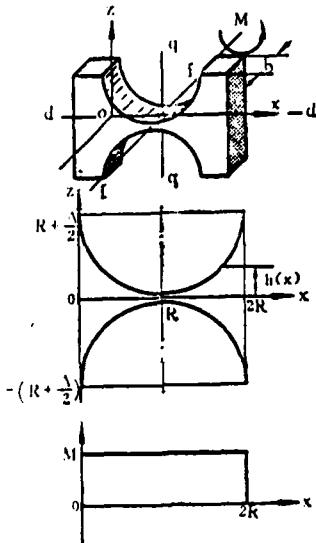


Figure I. The Calculation of the Angular Displacement Rigidity of the Axis of Bending

In this equation,  $R$  is the radius of the drilled aperture, and  $\Delta$  is the dimension of the thickness at the thinnest point of the component.

Let us assume that the angular deformation of the narrow neck component only occurs within the range of the radius of the drilled aperture. If we make this assumption, then, the total angular deformation of each single narrow neck component (angular displacement)  $\theta$  can be solved for by using the equation below, that is,

$$\theta = \int_0^{2R} \frac{M}{EI_s(x)} dx = \int_0^{2R} \frac{M}{E \cdot \frac{2}{3}bh^3(x)} dx \quad /140 \quad (I-3)$$

When the bending moment,  $M$ , is a constant value, then (I-3) can be written

$$\theta = \frac{3M}{2bE} \int_0^{2R} \frac{dx}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]^3} \quad (I-4)$$

If we use the angular displacement rigidity to express this, then, we have

$$K'_{\text{angular}} = \frac{2bE}{3 \int_0^{2R} \frac{dx}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]^3}} \quad (I-5)$$

The reason for this is that one must satisfy the requirement that the rigidities of the hinge assemblies be equal. Generally speaking, the widths of the narrow neck components of the interior sub-assembly and the exterior sub-assembly,  $b_{\text{interior}}$  and  $b_{\text{exterior}}$ , are not equal. Because of this fact, it is possible to write a representative equation for the angular displacement rigidity,  $K'_{\text{angle interior}}$ , of single narrow neck components of the interior sub-assembly. This equation is

$$K'_{\text{angle interior}} = \frac{2b_{\text{interior}} E}{3 \int_0^{2R} \frac{dx}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]^3}} \quad (I-6)$$

This matches the representative equation for the angular displacement rigidity,  $K'_{\text{angle exterior}}$ , of single narrow neck

components of the exterior sub-assembly. This equation is

$$\text{angle exterior} = \frac{2b_{\text{exterior}} E}{3 \int_0^{2\pi} \frac{dx}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]^3}} \quad (\text{I-7})$$

## II. THE CALCULATION OF VERTICAL AXIS OF NARROW NECK COMPONENTS UNDER THE INFLUENCE OF LOADING ALONG THE VERTICAL AXIS

Figure II shows a graphic representation of a single narrow neck component. Its structure is related to that shown in Figure I. In the same way, d-d, q-q, and f-f respectively are the vertical axis, cross axis and axis of bending of this narrow neck component.

If we assume that the load,  $P$ , comes along the direction of the vertical axis of the narrow neck component, then, the tensile deformation,  $d\delta$ , for the differential length,  $dz$ , in the direction influenced by the effects of the load can be calculated by using the equation below, that is,

$$\begin{aligned} d\delta &= \frac{Pdz}{Eb h_1(z)} \\ &= \frac{Pdz}{2Eb h(z)} \quad (\text{II-1}) \end{aligned}$$

In this equation,  $P$  is the load along the direction of the vertical axis of the narrow neck component.  $E$  is the modulus of elasticity of the material used, and  $b$  is the width of the narrow neck component along the direction of the axis of bending, which is a constant value.  $h_1(z)$  is the length of the narrow neck component along the direction of the  $x$  axis, and this length varies with changes in the value of  $z$ . Let  $h(z)$  be half the length, and we have  $h_1(z) = 2h(z)$ . In this case,  $h(z) = (R + \Delta/2) - \sqrt{R^2 - (z - R)^2}$ .

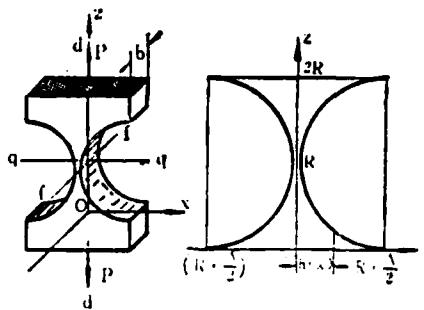


Figure II. The Calculation of the Tensile Rigidity in the Direction of the Vertical Axis.

When the load,  $P$ , is a constant value, and we recognize that tensile deformation is only produced within the scope of the radius of the drilled apertures, then, the total tensile deformation stress of a single narrow neck component is

$$\delta = \frac{P}{2Eb} \int_0^{2R} \frac{dz}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{R^2 - (z - R)^2} \right]} \quad (\text{II-2})$$

When we write this in the form of the vertical axis rigidity of the narrow neck component  $K'_{\text{vertical}}$ , then we have

$$K'_{\text{vertical}} = \frac{2Eb}{\int_0^{2R} \left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{R^2 - (z - R)^2} \right]} \quad (\text{II-3})$$

In equation (II-3), the width,  $b$ , is determined by variations in the calculated thickness of the component walls.

### III. THE CALCULATION OF THE CROSS AXIAL RIGIDITY OF A NARROW NECK COMPONENT UNDER THE INFLUENCES OF THE EFFECTS OF LOADING ALONG THE CROSS AXIS

Figure III shows an illustration of a single narrow neck component under the effects of loading along the cross axis. The structure of the narrow neck component is related to that shown in Figure I. In the same way as it is shown in Figure I, d-d, q-q, and f-f respectively define the vertical, cross and bending axes of a narrow neck component. It has already been pointed out in the main body of this article that, in fact, the loading on a narrow neck component is transmitted by one end of the component. Because of this, the load,  $P$ , along the line of the cross axis can make use of a moment of bending which has a numerical value equal to  $2PR$  (as is shown in the illustration, this is a negative value) and a moment of couple with a numerical value of  $PR$  (this is a positive value). These values may be used for purposes of substitution.

If one is speaking of the case of a single narrow neck component, then, the combined bending moment,  $M$ , which is exerting an influence on its surface, is expressed by the equation below, that is,

$$M(y) = P(R - y) \quad (\text{III-1})$$

Under the influence of this bending moment,  $M$ , the bending axis deformation,  $z$ , of the narrow neck component, can be solved for according to the equation for the elasticity curve. That is,

$$z'' = -\frac{M(y)}{EI_s(y)} \quad (\text{III-2})$$

In this equation,  $M(y)$  and  $E$  have clear meanings. As far as  $I_s(y)$  is concerned, it can be calculated on the basis of the equation below; this is possible according to the coordinate relationships illustrated in Figure III.

$$\begin{aligned}
 J_z(y) &= \frac{1}{12}b[2h(y)]^3 \\
 &= \frac{2}{3}bh^3(y) \\
 &= \frac{2}{3}b\left[\left(R + \frac{\Delta}{2}\right) - \sqrt{2Ry - y^2}\right]^3
 \end{aligned}
 \tag{III-3}$$

If we substitute this into equation (III-2), then, we can obtain

$$z'' = \frac{3P(R-y)}{2Eb\left[\left(R + \frac{\Delta}{2}\right) - \sqrt{2Ry - y^2}\right]^3}
 \tag{III-4}$$

If we do a double integration of (III-4), then, it is possible to solve and obtain the bending deformation along the z axis, that is to say,

$$\begin{aligned}
 z &= \iint \frac{3P(R-y)dydy}{2Eb\left[\left(R + \frac{\Delta}{2}\right) - \sqrt{2Ry - y^2}\right]^3} + Cy + D \\
 &= \frac{3P}{2Eb} \iint \frac{(R-y)dydy}{\left[\left(R + \frac{\Delta}{2}\right) - \sqrt{2Ry - y^2}\right]^3} + Cy + D
 \end{aligned}
 \tag{III-5}$$

Concerning the integration constants, in the case of C and D, these can be determined by the boundary conditions of the narrow neck components.

One should pay attention to the fact that the bending axis deformation which is related to the rigidity of the hinge assembly is the bending axis deformation along the axis in the place where  $y=R$  on the narrow neck component, that is to say,

$$K'_{\text{cross}} = \frac{P}{z_{(y=R)}} = \frac{2}{3} \left\{ \iint \frac{E b}{\left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Ry - y^2} \right]^3} + Cy + D \right\}_{(y=R)} - \quad (III-6)$$

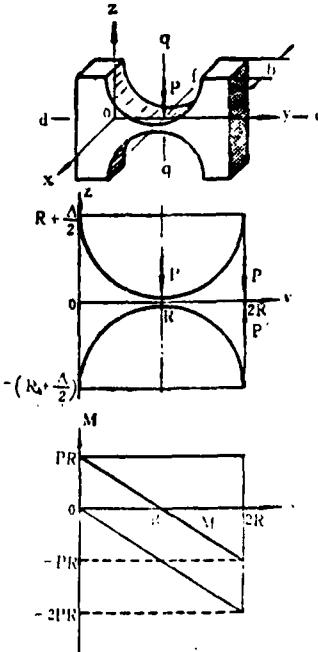


Figure III. The Calculation of the Cross Axial Rigidity Along the Direction of the Cross Axis.

In the same way, the width,  $b$ , in equation (III-6) varies with the calculated thickness of the component wall.

#### IV. THE CALCULATION OF THE CORRESPONDING RIGIDITY OF A NARROW NECK COMPONENT WHEN IT IS UNDER THE INFLUENCE OF LOADING ALONG ITS AXIS OF BENDING

Figure IV illustrates the situation when loading is applied along the axis of bending of a narrow neck component. All considerations relating to structural applications of narrow neck components in such situations are the same as those explained above. Due to the fact that the loading is also a

load,  $P$ , which is transmitted by one end of the narrow neck component under discussion and in the direction of the axis of bending, as was also the case in situations described above, it is possible to substitute the combined moment of bending formed from the moment of bending and the moment of couple. As far as the coordinate relationships involved in this illustration go, we have

$$M(x) = P(R - x) \quad (IV-1)$$

From the deformation of the axis of bending of the narrow neck component being considered,  $y$ , which is caused by the moment of bending  $M(x)$ , it is possible, according to the equation for the elasticity curve to make the calculations we are discussing. This equation is

$$y'' = \frac{M(x)}{EI_s(x)} \quad (IV-2)$$

In this equation, the cross section moment  $I_s(x)$  should be

$$I_s(x) = \frac{1}{12}[2h(x)]b^3 \quad (IV-3)$$

If we take  $h(x) = (R + \Delta/2) - \sqrt{2Rx - x^2}$  and substitute it into equation (IV-3), then, we can obtain

$$I_s(x) = \frac{1}{6}b^3 \left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right] \quad (IV-4)$$

If we take equation (IV-4) and substitute it into (IV-2), then, we obtain

$$y'' = \frac{6M(x)}{Eb^3 \left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]}$$

After integrating two times, it is then possible to obtain the bending deformation along the y axis, as shown below, that is,

$$y = \iint \frac{6P(R-x)dx dx}{Eb^3 \left[ \left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2} \right]} + C'x + D' \quad (IV-5)$$

In this equation, the integration constants,  $C'$  and  $D'$ , are determined from the boundary conditions of the narrow neck component being considered.

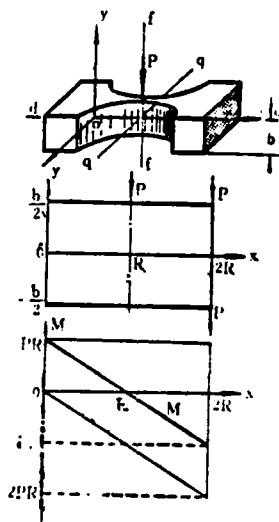


Figure IV. The Calculation of the Corresponding Rigidity With Loading Along the Axis of Bending.

In the same way, the rigidity calculated for the hinge assembly being considered is simply related to the bending deformation at the place where  $x=R$ , that is to say,  $y_{(x=R)}$ . Also corresponding to this idea is the fact that, when loading is added in the direction of the axis of bending, the rigidity calculated for a narrow neck component,  $K'$  bending, can be

figured on the basis of the equation below, that is

$$K'_{\text{bending}} = \frac{E b^3}{6 \left[ \iint \frac{(R-x)dx dx}{\left( R + \frac{\Delta}{2} \right) - \sqrt{2Rx - x^2}} + C'x + D' \right]_{x=0}} \quad (\text{IV-6})$$

#### REFERENCES

- [1] Howe, E. W. and Savet, P. H., The Dynamically Tuned Free Rotor Gyro, *Control Engineering*, Vol. 11, No. 6, 1964, pp. 67-72.
- [2] Craig, R. J. G., Theory of Operation of an Elastically Supported Tuned Gyroscope, *I. E. E. E. Trans. on Aerospace and Electronic Systems*, Vol. AES-8, No. 3, May. 1972, pp. 280-288.
- [3] Craig, R. J. G., Theory of Errors of a Multigimbal, Elastically Supported Tuned Gyroscope, *I. E. E. E. Trans. on Aerospace and Electronic Systems*, Vol. AES-8, No. 3, May 1972, pp. 289-297.
- [4] Ensinger, W. B. and Spring, M. S., Gyroscope Flexure Hinge Assembly, U. S. Patent No. 3, 614, 894, Oct. 1971.
- [5] M.M (name which sounds like Fei Lc Ning Ke-Bao Lo Di Qi (ed): *Material Dynamics Course*, Gao Deng Jiao Yu Chu Ban She, 1956, Chinese language edition.

A THEORETICAL ANALYSIS OF INVOLUTE HARMONIC GEARING

Shen Yunwen

A Translation of "Jian Kai Xian Xie Bo Chi Lun Zhuan Dong De  
Yan Jiu", from Xi Bei Gong Ye Da Xue Lun Wen Xuan (Part I),

1979

## Summary

# A Theoretical Analysis of Involute Harmonic Gearing

*Shen Yunwen*

The harmonic gearing is a new type drive which is being developed in China and abroad. Thus far, no theoretical proof has been given to justify the use of involute profile. The known methods for calculating transmission error and for the analysis of power harmonic gearing are inadequate in completeness. A method for selecting the optimum parameters of engagement with given backlash has not yet been obtained. It is the purpose of this paper to fill in some gaps in the theory of harmonic gearing.

This paper studies the following four main problems:

First, starting from the study on engagement theory of harmonic gearing, this paper derives an equation of the theoretical profile and gives its numerical solution. The error of the technological profile (involute profile) as compared with the theoretical profile is evaluated by means of optimum approximation. The results of calculation show that the maximum profile error is about 0.0016 m. Therefore, it has been strictly proved appropriate to substitute involute profile for theoretical profile.

Secondly, this paper discusses the effect of involute profile on transmission error. On the basis of internal gearing with involute profile, the formulae for calculating the instantaneous velocity ratio in normal and border contact are provided.

Thirdly, the author presents the backlash control method and non-interference criterion which are needed in optimum engagement design. The backlash control method is essentially a constrained optimization method satisfying certain constraining relationships (non-interference, meshing depth of the tooth  $(r_{a1} + w_1) - r_{a2} \geq m$ , addendum thickness  $S_a \geq 0.25$  m, radial clearance  $r_{f2} - (r_{a1} + w_1) \geq 0.2$  m, and others), the minimum of the objective function  $f(\mathbf{x}) = j_e(\mathbf{x}) - j_e$  being found by Hooke-Jeeves' method. As a result of optimum design, the geometrical parameters  $\xi_1, \xi_2, h_a$  are obtained for a given minimum backlash  $j_e$ .

Finally, by using a discrete model and considering the distortion of the flexspline, a method for analyzing power harmonic gearing is provided. The requirements for contact between elements of the harmonic gearing are described by equations  $w_{ek} - w_k = \lambda_{uk} F_k$  and  $j_{uk} = -\lambda_{uk} T_k$ . Using these equations,

we find the forces acting on the meshing teeth and the load distribution along the arc of action of flexspline, and then obtain the actual deflected middle line of flexspline. Other interesting problems such as replacing actual loading tests of harmonic gearing with computer analysis, modification of the cam profile of generator, and backlash selection, may be solved by means of this method.

# A THEORETICAL ANALYSIS OF INVOLUTE HARMONIC GEARING

Shen Yunwen

## SUMMARY

This article begins with a study of the influences of errors in the theoretical profiles of gears, as well as in their technological profile of operation, and also initially considers the influences of errors in transmission. It makes a rigorous theoretical demonstration of the feasibility of using involute gearing profiles to replace theoretical gearing profiles. The article presents conditions in which transmission will not give rise to interference and uses optimization techniques as part of a practical method for solving the problem of controlling side cracking. At the end of this article, in considering situations involving the deformation of soft gears, it analyses several questions dealing with dynamic harmonic gearing transmission, and presents ways of solving these problems.

From the time when C.W. Musser [1] first coined the name harmonic transmission for this new form of transmission, there has been large scale research work carried out by both China and other countries into this type of transmission, and this research has achieved considerable success. At present, in the various financial agencies of our government, this type of transmission is undergoing an increased expansion of its use, and an even deeper research effort is going into the use of involute gear types in harmonic gearing transmission and the practical significance that this combination will have.

This article is going to discuss several important questions which need solving in the areas of design and research involving the use of involute harmonic gearing in transmissions. The article will initially begin its study of this topic with research into the theoretical gearing profiles for a trans-

mission involving harmonic gearing. After numerical values are solved, in order to describe the theoretical gearing profile, then, the article makes use of the best approximation method to make a detailed discussion of the errors related to the use of involute forms in operating gear profiles. This discussion goes a step further in explaining the influence which this type of operating gear profile has on transmission errors. The article then goes on from there to proving, by the use of theory and practical applications, both the rationality and feasibility of using involute gearing profiles as the operating gear profiles for harmonic gearing. Following this, this article discusses practical problems which need to be solved in regard to the analysis of the gear engagement in this type of transmission. This section includes such problems as the conditions under which interference does not occur, methods for the control of backlash, and other related problems. Moreover, in the engineering aspect of the problem, this article presents a practical design method for making use of the optimized method for carrying out the selection of the best possible gear meshing parameters. And, besides this, this article offers routes toward the solution of problems involved in the investigation of the loading distribution between gear teeth in dynamic harmonic gear transmissions, the adjustment of the gear profile lines in wave generators, and the selection of backlash values and other related problems of a similar type. In the final section of this article, it also presents some routes along which to reach solutions of the problems associated with the methods of analysis used when dealing with elastic deformations in the original curves of flexspines.

#### I. RESEARCH INTO THEORETICAL GEAR PROFILES AND OPERATIONAL GEAR PROFILES AS WELL AS THEIR NUMERICAL ANALYSIS

As is commonly known, theoretical gear profiles in harmonic gear transmission are, in general, very difficult to realize

technically. Moreover, it is not possible to use elementary methods in order to describe the mathematical curves involved. In order to facilitate the manufacture and testing of this sort of gearing, it is necessary to make use of a type of gearing profile which is easier to realize from a technical point of view and use this profile as the technological profile to replace the theoretical profile we have mentioned. What this article discusses is the type of transmission which has harmonic gearing and makes use of involute curves as the profiles for this gearing; moreover, this article also gives the name involute harmonic gear transmissions to this type of set up.

In order to demonstrate the reasonableness and feasibility of using involute gearing profile curves as the technological profiles of gearing, it is first necessary to study the theoretical gearing profiles of harmonic gear transmissions. After this, one must then analyze the differences between the technological profiles involved and the corresponding theoretical profiles. The theoretical gearing profiles of transmissions with harmonic gearing point to the use of envelope theory from classical geometry in order to solve the conjugate gearing involved. During the process of reaching a solution for this conjugate gearing, it is possible to take the gear profile of a certain gear wheel, solve for it, and then, use that result to reach a solution for the conjugate of yet another gear wheel gearing profile.

If one is considering the case in which the gearing profile of a flexspine is an involute curve, then, in the case in which the form of a wave generator is already selected (corresponding to the original curve,  $\mathcal{O}$ , which is already known), it is easy to solve for the gearing profile  $G$  of a rigid gear wheel.

When one is making this sort of discussion, the assumptions involved are as follows:

1. In the process of transmission, the length of the centerline of the flexspine involved does not change.
2. The moment of inertia of the vertical cross section of a flexspine, must be much larger than the moment of inertia of the vertical cross section through the area between flexspines. When this is the case, it is possible to note the fact that, during operation, flexspines do not undergo deformations. It is only the space between them which deforms.
3. Flexspines are cut out in situations in which there is no deformation involved.
4. Under the influences of the forces of deformation and the forces impelling the gears to mesh, the configuration of the curve of elastic deformation for the centerline of flexspines is stable and non-deformed. It is also assumed that we are ignoring the tiny vibrations which take place around the average positions which correspond to the elastic curvature. This is caused by the fact that, when the gears mesh, the placement of the points of contact is not the same every time the gears involved go around.

In order not to lose the universal character of the problems being discussed, this article also discusses the case which involves a fixed wave generator.

As is shown in Figure 1 it is assumed that there is a fixed coordinate system  $\{X_0Y_0\}$  and a wave generator which is fixed in this system. The origin, 0, is located in the center of the wave generator, and the Y axis is congruent with the long axis of this generator. There is also a dynamic coordinate

system  $\{x_1, o_1, y_1\}$  in which flexspines are set. The origin in this case,  $o_1$ , is located at a certain point,  $c$ , on the original curve,  $\sigma$ . The  $y_1$  axis is congruent with the axis of symmetry of the flexspines involved. In yet another dynamic coordinate system,  $\{x_2, o_2, y_2\}$ , there are rigid gears set.  $o_2$  is located in the center of rotation of the rigid gears involved. The  $y_2$  axis is congruent with the axis of symmetry between the gear teeth. Because of this, when a flexspine gear has an angular velocity  $\omega_1$  and is turning in the direction shown in the Figure, then, the teeth of this flexspine gear will move along the surface of the wave generator. Moreover, this same force will propel the teeth of rigid gear wheels in their movements as well. At such times, as far as flexspines are concerned, except for the movements which follow along together with the  $c$  point on the original curve,  $\sigma$ , the axis of symmetry of these gears will still correspond to the rotation of point  $c$  through an angle  $\mu$ . Because of this, the origin of the  $\{x_1, o_1, y_1\}$  system,  $o_1$ , (that is, the  $c$  point) has coordinates which, in the  $(\rho, \varphi_1)$  system, are capable of being represented in the form of a first degree approximation as

$$\begin{cases} \rho \approx r_m + w, \\ \varphi_1 \approx \varphi + \frac{v}{r_m}. \end{cases} \quad (1)$$

In these equations,  $\rho$  is the polar radius of the original curve in a polar coordinate system.

$r_m$  is the radius of the center line of a flexspine before deformation takes place.

$w, v$  are, respectively, the symbols which represent the radial displacement and shear displacement of points on the centerline of flexspines after deformation takes place. When one is considering the case in which use is made of the deformation curvature of the centerline of the horizontal cross

section of the column casing at its two ends, when under the influence of the four forces which deform the long axis from an angle  $\beta$ , then, on the basis of (2), it is not difficult to solve for  $w$  and  $v$  using the equations shown below, that is,

$$\left\{ \begin{array}{l} w = \frac{w_0}{\sum_{n=2,4,6,\dots} \frac{\cos n\beta}{(n^2-1)^2}} - \sum_{n=2,4,6,\dots} \frac{\cos n\beta \cos n\varphi}{(n^2-1)^2}, \\ v = - \frac{w_0}{\sum_{n=2,4,6,\dots} \frac{\cos n\beta}{(n^2-1)^2}} \sum_{n=2,4,6,\dots} \frac{\cos n\beta \sin n\varphi}{n(n^2-1)^2}. \end{array} \right. \quad (2)$$

In these equations,  $w_0$  - is the maximum radial displacement.

$\varphi$  - is the angle of rotation of the non-deformed end of a flexspine which is necessary to make the angle of incidence into correct angle for the situation. Concerning  $\varphi = i_0 \varphi_2$ , and  $i_0 = \frac{z_2}{z_1}$ ,

$\varphi_2$  is the angle of rotation of a rigid gear wheel, and  $z_1$ ,  $z_2$  are respectively the number of teeth in flexspines and rigid gears.

From (3), one obtains the equation for representing the gear profile,  $\bar{x}$  for the right side of a flexspine in the coordinate system  $(x_1, y_1)$ , that is

$$\left\{ \begin{array}{l} x_1 = r_1 [-\sin(u_1 - \theta_1) + u_1 \cos \alpha \cos(u_1 - \theta_1 + \alpha)], \\ y_1 = r_1 [\cos(u_1 - \theta_1) + u_1 \cos \alpha \sin(u_1 - \theta_1 + \alpha)] - r_m. \end{array} \right. \quad (3)$$

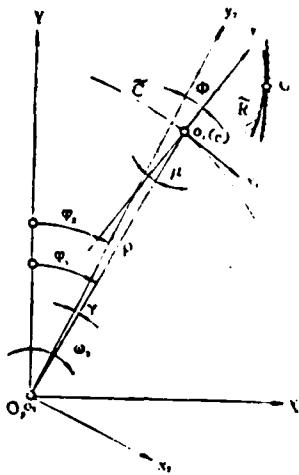


Figure 1

In these equations,  $r_1$  - is the graduation of the circular radius of flexspines

$u_1$  - is a parameter representing the angle of rotation which is determined from the conditions which prevail between a cutting tool and the flexspine being cut when there is no slippage or rolling of the flexspine being cut.

$\alpha$  - is the original gear profile angle of the cutting tool involved.

$\theta_1$  - is half the angular thickness of a gear tool on a round gear, that is,  $\theta_1 = \frac{1}{2r_1} \left( \frac{\pi}{2} + \alpha_1 \right) m$ . In this equation,  $\alpha_1$  is parameter representing the change in the angular thickness of a gear tooth on a round gear. It is possible to do calculations on the basis of the meshing

displacement of a gear, and in this case  $m$  is, then, a modulus for this.

Due to all of this, on the basis of envelope theory, it is possible to solve for an equation to represent the gearing profile,  $G$ , for a rigid gear which is the conjugate of the gearing profile,  $R$ , of a flexspine, that is,

$$\begin{aligned}
 x_2 &= x_1 \cos \phi + y_1 \sin \phi + \rho \sin \gamma, \\
 y_2 &= -x_1 \sin \phi + y_1 \cos \phi + \rho \cos \gamma, \\
 \frac{\partial x_2}{\partial u_1} \frac{\partial y_2}{\partial \varphi} - \frac{\partial x_2}{\partial \varphi} \frac{\partial y_2}{\partial u_1} &= 0, \\
 \phi &= \gamma + \mu, \\
 \gamma &= \varphi_1 - \varphi_2, \\
 \mu &\approx \frac{w_0}{r_m} \sum_{n=2,4,6,\dots} \frac{\cos n\beta}{(n^2 - 1)^2} \sum_{n=2,4,6,\dots} \frac{n \cos n\beta \sin n\varphi}{(n^2 - 1)^2}.
 \end{aligned} \tag{4}$$

In these equations, the angles of rotation,  $\phi_1$  and  $\phi_2$ , both represent the angles of rotation shown in the Figure as they relate to the base or undeformed direction.  $\mu$  represents the angle of rotation which is necessary to make the angle of incidence coincide with the undeformed base direction for a radius  $\rho$ .  $\phi$  is the included angle needed in order to represent the two coordinate systems  $\{x_1, y_1\}$  and  $\{x_2, y_2\}$  relative to each other, and  $\varphi_1, \varphi_2, \mu, \rho$  are all functions of  $\phi$ .

If we take equation (3) and substitute it into equation (4), then, the equation for the theoretical profile for the gearing of a rigid gear wheel can be written in the form

$$\begin{cases}
 x_1 = r_1[\sin \zeta + u_1 \cos \alpha \cos \lambda] - r_m \sin \phi + \rho \sin \gamma, \\
 y_1 = r_1[\cos \zeta - u_1 \cos \alpha \sin \lambda] - r_m \cos \phi + \rho \cos \gamma, \\
 r_1 \dot{\phi}[\cos \alpha \sin \alpha + u_1^2 \cos^2 \alpha] - r_m \dot{\phi}[\sin(\zeta - \phi) \\
 + \cos \alpha \sin(\phi - \lambda) + u_1 \cos \alpha \cos(\phi - \lambda)] \\
 - \rho[\cos \alpha \cos(\lambda - \gamma) - \cos(\zeta - \gamma) + u_1 \cos \alpha \sin(\lambda - \gamma)] \\
 - \rho \dot{\gamma}[\cos \alpha \sin(\lambda - \gamma) - \sin(\zeta - \gamma) - u_1 \cos \alpha \cos(\lambda - \gamma)] = 0, \\
 \zeta = \phi - (u_1 - \theta_1), \\
 \lambda = \phi - (u_1 - \theta_1 + \alpha).
 \end{cases}
 \quad (5)$$

In these equations  $\phi, \dot{\gamma}, \rho$  are all derived from  $\phi, \gamma, \rho$  as they relate to  $\phi$ .

Equation (5) can only be used to obtain numerical solutions. When solving, one must first solve for the value of  $\phi$  which corresponds to  $u_1$  in the third equation among those shown above. Then, one must solve for  $\gamma, \phi, \rho$  as well as the corresponding values,  $\zeta, \lambda$ . After substituting the first and second equations among those in (5), it is easily possible to solve for the coordinate values for the gearing profile for a rigid gear wheel,  $G$ .

It is not difficult to see that the third equation among those shown in (5) is an equation which represents a function, the form of which is  $F(u_1, \phi) = 0$ . It is only necessary, within the range of values which satisfy the conditions  $u_{f1} \leq u_1 \leq u_{al}$  ( $u_{f1}$  and  $u_{al}$  are respectively the parameters which correspond to the gearing profile for a flexspine in its base section and its tip section) to select points with appropriate numerical values, and then it is easy to make use of the Muller method to solve for the values  $\phi$  which correspond to these points.

Concerning harmonic gear transmissions with the parameters  $z_1=200$ ,  $z_2=202$ ,  $\alpha=20^\circ$ ,  $m=0.5$ , when the form of the original curve is influenced by the four forces which pertain when  $\beta=30^\circ$ , and one selects  $w_0=m=0.5$ , wall thickness  $\delta=0.9$ ,  $r_m=50.375$  mm, and the coefficient of flexspine displacement  $\xi_1=3.0$ , then one can solve for the numerical values of the theoretical gearing profile of a rigid gear. One can use the curve in Figure 2 to represent this form.

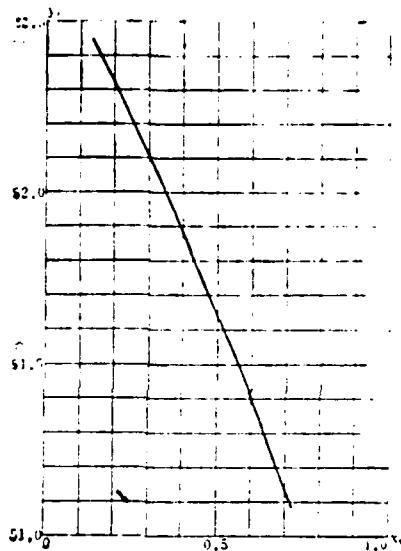


Figure 2

In order to make it more convenient to use involute gearing forms as technological gearing profiles, it is necessary to use involute gearing profiles to make an optimum approximation to the theoretical gearing profile of rigid gears and in order to settle on reasonable meshing parameters. In making this approximation, in order to guarantee that the gearing profiles involved will not give rise to interference, it is necessary to require that involute gearing profile curves involved be to the right of the curves for the theoretical gearing profiles. Moreover, it is also necessary to require that the average values involved in the computations of the distances between the

corresponding points on the involute curve gearing profiles and the theoretical gearing profiles act as a control on errors in the process of approximation, that is to say, that one should cause the average error

$$\epsilon = \frac{\sum d_k}{n}$$

to be as small as possible. The quantity,  $d_k$ , in this equation is the distance between the  $k$ th corresponding points on two gearing profiles. Obviously,  $d_k = f(\xi_2, u_{2k})$ .  $\xi_2, u_{2k}$  are, respectively, the parameters for the displacement and the  $k$ th point on the involute gearing profile of a rigid gear wheel. Because of this,  $\epsilon$  is also a function of  $\xi_2$  and  $u_{2k}$ . Concerning the necessity of satisfying the condition that  $\epsilon$  be as small as possible and  $d_k \geq 0$ , this is a question of solving for values corresponding to the extremes of the conditions involved. It is possible to make use of a one-dimensional search method in order to carry out a unilateral approximation in a positive direction. The results of calculations by computer show that, when  $\xi_2 = 2.66696$ , the best results are obtained. At this time  $\epsilon = 0.00053718$  mm. The largest error involved in this is 0.0007723mm. This corresponds to 0.0016m. Moreover, in Figure 2, the curves for involute gearing profiles and theoretical gearing profiles are already congruent.

From this, one can see that it is only necessary to make a reasonable selection for the displacement parameter involved, and the involute gear form curve is capable of very closely approximating the theoretical gearing profile form. Moreover, the manufacturing processes involved in such a case are easily managed, and it is possible to make good use of gear wheel cutting tools and testing equipment which is available "off-the shelf." Because of this, it is possible to make use of involute

AD-A104 327

FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OH  
RECENT SELECTED PAPERS OF NORTHWESTERN POLYTECHNICAL UNIVERSITY—ETC(U)

F/6 20/4

AUG 81

UNCLASSIFIED FTD-ID(RS)T-0259-81-FT-1

NL

405  
AD  
AD-104 327

EX-100  
DRAFT  
04-811  
DTIC

CONT

curve gear forms as the technological gear profiles of harmonic gears. This application is also reasonable, and the method is worthy of expanded use. It is also true that other possible shapes for gear teeth (triangular teeth, for example) lack the excellent characteristics of these harmonic forms.

The forms of analysis which have been discussed above also lead directly to a type of method for the determination of displacement parameters for harmonic gear transmissions when there is no backlash present in the meshing of the gears.

## II. THE INFLUENCE OF INVOLUTE GEARING PROFILES ON THE PRECISION OF HARMONIC GEAR TRANSMISSIONS

When flexspines and rigid gear wheels both make use of involute gear profiles, then, if their parameters of displacement are solved accurately on the basis of the methods discussed above, it is possible to recognize the fact that the profiles of the two gear wheels are conjugates, and that both of them are capable of realizing a condition in which there is no backlash in meshing of the gears. However, practically speaking, if one makes use of transmissions with differing requirements, then, there will be different requirements in terms of backlash. Because of this, it is necessary to make the selection of meshing parameters on the basis of these different requirements. In such a situation, the coefficients of displacement which are chosen will not necessarily be a match for the requirements set by the optimum approximation of the involute gearing profile involved to the corresponding theoretical gearing profile. Because of this fact, the two types of gear wheel profiles involved in a situation like this are approximate conjugates, and, because of this, it is not possible to guarantee that there will be precise rules governing the movements involved in the transmission. The rates of instantaneous

transmission in a situation like this will produce changes. Concerning the rules which govern changes in the rates of instantaneous power transmission in harmonic gearing, these rules depend not only on the configuration of the outline profile of the wave generators involved but also on the configuration of the gearing profiles. What this section will discuss is nothing else but to try to present a universally applicable type of method. With this method, (after one has a given original curve, and one is then confronted with unexpected errors in the manufacture and installation of a system,) one can make a precise determination of the errors in transmission caused by the approximate character of the gearing profile curves involved.

Concerning the question of flexspines as flexible components, when a flexspine follows a course of movement along the surface of a stationary wave generator, then on the basis of the assumption that, after deformation takes place, each cross section of the flexspine returns to maintaining its original plane form, it is possible to recognize the fact that the instantaneous center of revolution for each cross section of a flexspine body is distributed somewhere along the involute curve,  $\mathfrak{S}$ , of the original curve (Figure 3). From the Willis principle, we know that, if the point of contact between two gear profiles,  $R$  and  $G$ , is  $k$ , then, from the figure one can know that the point of intersection,  $P$ , of the curve through  $k$  and the extension of the curve formed by connecting the centers of instantaneous revolution into the line  $OO$ , is nothing but the nodal point of instantaneous meshing. Because of this fact, the ratio of instantaneous transmission for this meshing position is

/149

$$i_m = \frac{PO}{PO_0} = \frac{PO}{PO - a}.$$

(6)

The value of  $a_c = \infty$  can be solved for without great difficulty through differential geometry. If the coordinates of  $O_a$  are  $(X_a, Y_a)$ , then,

$$\left\{ \begin{array}{l} a_c = \sqrt{X_a^2 + Y_a^2}, \\ X_a = \rho \sin \varphi_1 - \frac{(\rho^2 + \rho^2)(\rho \sin \varphi_1 + \rho \cos \varphi_1)}{\rho^2 + 2\rho^2 - \rho \dot{\rho}}, \\ Y_a = \rho \cos \varphi_1 - \frac{(\rho^2 + \rho^2)(\rho \cos \varphi_1 - \rho \sin \varphi_1)}{\rho^2 + 2\rho^2 - \rho \dot{\rho}}. \end{array} \right. \quad (7)$$

The value of  $PO$  can be solved for on the basis of the geometrical relationships which relate to the involute curves present in the meshing of gear transmissions.

Due to the fact that gear wheels are machined in a state before the application of any deforming forces, if we assume that, during operation, there is no deformation of these gear wheels, then, due to these facts, when flexspines are deformed, their base circles will move following along with the gear wheels involved on the basis of fixed relationships which are maintained between gear wheels throughout the movement. From this perspective, it is possible to view transmission by harmonic gearing as a series of gear wheels with different center distances meshing with rigid gear separately at different meshing positions. Because of the fact that the center of base circles is supposed to be located on the curve of symmetry of a gear wheel, at the meshing positions shown in the illustrations, the centers of the base circles are located at  $O_k$ . At such a time, the meshing angle  $\alpha'_k$ , which corresponds to the particular meshing position  $k$ , can be solved for according to the normal methods. For the case of double wave transmissions, this means

$$\alpha'_k = \arccos \frac{(z_2 - z_1)m \cos \alpha}{2a'_c} = \arccos \frac{m \cos \alpha}{a'_c}, \quad (8)$$

In this equation

$$a'_c = \overline{OO_{11}} = \sqrt{r_m^2 + \rho^2 - 2r_m\rho \cos\mu} \quad (9)$$

The geometrical significance of  $a'_c$  and  $a'_c$  is obvious. However, in the case of practical transmissions, the center of instantaneous revolution for the meshing positions concerned is not  $O_{11}$ . It is  $O$ . Because of this fact, the actual angle of meshing in such a case should be  $\alpha_k = a'_c + \Delta a_k$ .  $\Delta a_k$  can be solved for by the use of cosine theorems  $\Delta O_1 O_{11}$ . If we assume that the radius of curvature at point c on the original curve is  $r_c = \overline{O_c c}$ , then,  $r_c$  can be solved for the radius of curvature equations used to solve for curves in differential geometry. It is also possible to reach solutions by making direct use of the coordinates  $O_1(X_1, Y_1)$  and  $c(X_c, Y_c)$ . When this method is used,  $\overline{O_1 O_{11}} = r_c - r_m$ , and, on the basis of this

$$\Delta a_k = \arccos \frac{a_c^2 + a_c'^2 - (r_c - r_m)^2}{2a_c a_c'} \quad (10)$$

When  $a'_c > a_c$ ,  $\Delta a_k$  takes on a negative value. In the opposite case, it takes on a positive value. Because of these facts,

$$\begin{cases} \overline{PO} = m z_2 \frac{\cos\alpha}{\cos\alpha_k}, \\ \alpha_k = a'_c + \Delta a_k. \end{cases} \quad (11)$$

If one is considering the case in which equation (7) and equation (11) are substituted into equation (6), that is to say that it is possible to solve for the instantaneous transmission ratio,  $i_M$ , for a given meshing position, then, on the basis of this approach, the amount of change in the instantaneous transmission ratio is

$$\Delta i = i_M - i_0. \quad (12)$$

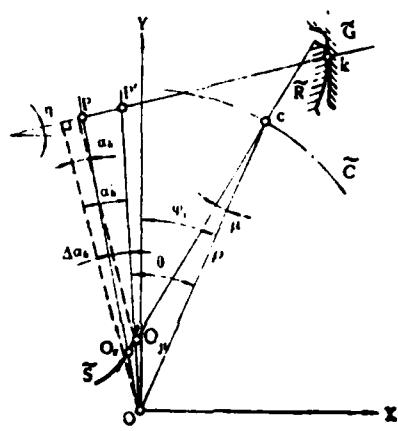


Figure 3

The magnitude of  $J$  verifies the magnitude of the transmission error. However, the methods of calculation discussed above are only suitable for use inside the limits of the normal range of meshing. In the same way, the point of contact,  $k$  should be located within the operational section of a gear profile. If the point of contact,  $k$  is located outside the technological range of a gear profile, then, theoretically speaking, there ought to be no contact. If we follow this line of reasoning in the direction, then, due to the results of the elastic properties of flexspines, even if the gears do come in contact with each other, the gear teeth will then be put into a configuration in which only the tips of the teeth are in contact.

The condition which decides whether or not gear wheels are in a normal meshing configuration is

$$r_{a2} \leq \sqrt{X_k^2 + Y_k^2} \leq r_{g2} \quad (13)$$

In this relationship,  $r_{a2}$  - is the apical radius of a rigid gear  
 $r_{g2}$  - is the radius at the end point of an  
involute curve on a rigid gear wheel  
 $X_k, Y_k$  - the coordinates of contact points  
within  $\{XOY\}$

The coordinates  $(X_k, Y_k)$  are determined by the intersection point of the common normal with the gearing profile of the rigid gear wheel involved. In order to simplify the calculations, the gearing profile for the rigid gear wheel can be replaced by a straight line (the error in this case does not exceed 0.01m). The equation for the normal line which crosses the point of contact is

$$\begin{cases} Y = K_1 X + \frac{r_2 \cos \alpha}{\cos \eta}, \\ K_1 = \tan \eta, \\ \eta = \theta + \alpha'_k - \varphi_{13}, \\ \theta = \arccos \frac{\alpha'^2_k + \rho^2 - r_m^2}{2\alpha'_k \rho}. \end{cases} \quad (14)$$

The equations which represent the rigid gear gearing profile within  $\{XOY\}$  which correspond to the meshing positions which were studied are

$$\left\{ \begin{array}{l} Y - Y_{a2} = K_2(X - X_{a2}), \\ K_2 = \frac{Y_{r2} - Y_{a2}}{X_{r2} - X_{a2}}, \\ X_{a2} = r_2 \{ \sin[\varphi_2 - (u_{a2} - \theta_2)] + u_{a2} \cos \alpha \cos[\varphi_2 - (u_{a2} - \theta_2 + \alpha)] \}, \\ Y_{a2} = r_2 \{ \cos[\varphi_2 - (u_{a2} - \theta_2)] - u_{a2} \cos \alpha \sin[\varphi_2 - (u_{a2} - \theta_2 + \alpha)] \}, \\ X_{r2} = r_2 \{ \sin[\varphi_2 - (u_{r2} - \theta_2)] + u_{r2} \cos \alpha \cos[\varphi_2 - (u_{r2} - \theta_2 + \alpha)] \}, \\ Y_{r2} = r_2 \{ \cos[\varphi_2 - (u_{r2} - \theta_2)] - u_{r2} \cos \alpha \sin[\varphi_2 - (u_{r2} - \theta_2 + \alpha)] \}. \end{array} \right. \quad (15)$$

In these equations  $X_{a2}$ ,  $Y_{a2}$  and  $X_{r2}$ ,  $Y_{r2}$  are respectively the coordinates for the rigid gear apex and gear base with  $\{XOY\}$ . The meaning of the other symbols is the same as it has been in other equations before. The subscript "2" under some symbols stands for the case of a rigid gear. However, one must take careful note of the fact that  $\theta_2$  is calculated from the interval between rigid gears.

On the basis of all of this, the coordinate values which are obtained for the point of contact are

$$\left\{ \begin{array}{l} X_b = \frac{K_2 X_{a2} - Y_{a2} + \frac{r_2 \cos \alpha}{\cos \eta}}{K_2 - K_1}, \\ Y_b = \frac{K_1 (K_2 X_{a2} - Y_{a2}) + K_2 \frac{r_2 \cos \alpha}{\cos \eta}}{K_2 - K_1}. \end{array} \right. \quad (16)$$

The situation in which one is dealing with contact only on the tips of the gear teeth is relatively complicated. Because of this, the instantaneous transmission ratio,  $i_{mb}$ , for such a case can only be calculated in the form of an estimate.

In the case in which there is a fixed wave generator, if one recognizes in an approximate way that the polar radius  $r_{\alpha}$  which exists as an instantaneous quantity at the top of the flexspines being studied does not change, then, when the in-coming end of a flexspine turns through  $d\varphi$  degrees of angle, the minute angular displacement through which the flexspine turns is

$$d\varphi_{a1} = d\varphi + \left[ \frac{\dot{\psi}}{r_m} + \frac{(r_{a1} - r_m)\dot{\alpha} + \dot{w} \operatorname{tg} \alpha_{a1}}{r_{a1}} \right] d\varphi,$$

If at this time, the angle through which the rigid gear wheel turns,  $d\varphi_1$ , satisfies the condition of being equal to the rotational angle of the point of contact concerned, then, one obtains

$$i_{m1} = \frac{r_{a1}}{r_{a1} \left( 1 + \frac{\dot{\psi}}{r_m} \right) - (r_{a1} - r_m)\dot{\alpha} + \dot{w} \operatorname{tg} \alpha_{a1}} \quad (17)$$

In this equation,  $\alpha_{a1}$  is the gear profile angle for a flexspine, and it is possible to make an approximate solution for it on the basis of the straight line gear profile.  $\dot{w}, \dot{\psi}$  and  $\dot{\alpha}$  both are derivatives of  $w, \psi, \alpha$  from  $\phi$ . The polar radius which exists for the top of a flexspine at a given meshing position should theoretically be determined by the coordinates for the top of the gear at that point of meshing being considered; however, because of the fact that  $w$  is very small (usually less than  $2^\circ$ ), it is possible, in order to simplify calculations, to get a solution of an approximate sort by using the equation below

$$r_{a1} \approx r_{a1} + w. \quad (18)$$

In this equation,  $r_{a1}$  is the apical circle radius of a flexspine. If one solves for  $i_{m1}$ , then, it is possible to solve for  $\dot{\alpha}$ .

Calculations demonstrate that, as far as the parameters of transmission which were discussed before go, when we take  $\xi_1=3.075$  and  $\xi_2=2.746$ , the normal scope of the area of meshing is approximately  $4.75^\circ$ . The error in the instantaneous transmission ratio, which is caused by the approximate nature of the form of the gears involved is not larger than 0.5%. It can be seen that, because of the fact that the influence which this exerts on the precision of transmission is not large, during the design process, there is no necessity to go back through the procedures which were used before, that is to say, solving for the numerical values of a theoretical gearing profile and then using a numerical approximation method in order to determine the coefficients of displacement, and so on. Instead of this, it is possible to make use of direct computer selection of the parameters of displacement on the basis of the backlash requirements involved. This provides data for research into the control of backlash.

Because of the fact that the transmission of power by harmonic gearing has a form of meshing of gears which is complicated in the extreme, in the field of theoretical calculations, it is usually only possible to make a qualitative estimate and it is difficult to completely reflect the actual situation. When one is considering the case in which the effects of moments of bending are involved, deformations due to elastic twisting in flexspines as well as gear wheel deformation cause a simultaneous increase in the number of teeth meshing, and each of these influences has its own independent ratio of instantaneous transmission relating to the gear teeth which are meshing. At the same time, due to the fact that gear contacting at acute angles have relatively small rigidities as compared to the total rigidity of these same gears in different circumstances, most of the transmission errors for these gears is averaged out by the effects of elastic deformation. Moreover, the elastic deformation of components carries with the capability of leading to an enlargement of the normal area of

meshing (when the bending moments involved are large enough to cause the appearance of distortions in the original curves involved, then, it will be necessary to do other analyses). This type of overall averaging effect which is caused by elastic deformations is one of the reasons for the high precision of power transmission by harmonic gearing.

### III. MESHING INTERFERENCE PROBLEMS AND BACKLASH CONTROL

Among the phenomena associated with interference in harmonic transmissions, the main ones are the following: superimposition interference in gearing profiles and excessive curvature interference.

If one wants to eliminate the occurrence of interference when two gears are meshing with each other, then, it is necessary, at a given point of meshing, for the circumferential spacing to be  $i \geq 0$ . Because of this, interference calculations and the control of the amount of backlash are two aspects of the same problem. This problem is how, during the design of a transmission, one is to maintain, within the conditions necessary for non-interference, a backlash which can be obtained at a reasonable value.

#### 1. The Conditions for Preventing the Occurrence of Interference in the Instants when Gears are Meshing and Unmeshing

Obviously, the condition for the non-occurrence of gearing profile superimposition interference at the moments when the gears have just meshed is as follows: the angle of rotation  $\phi_L$  which corresponds to the coordinates for the top of the gear teeth on the technological tooth surface of a rigid gear should be larger than the angle of rotation  $\phi_L$  which corresponds to the coordinates for the top of the teeth on the technological gear tooth surface of a flexspline.

Due to the fact that at the instant when gears are meshing (or unmeshing), in  $\{XOY\}$  the coordinates for the tops of flexspline teeth are equal to the coordinates for the tops of rigid gear teeth in contact with them. Because of this

$$\sqrt{X_{a_1}^2 + Y_{a_1}^2} = r_{a_1}. \quad (19)$$

If we take the equation which represents the flexspline gearing profile in  $\{XOY\}$  and substitute it into equation (19), and we go through the appropriate transformations, then,

$$x_{a_1}^2 + y_{a_1}^2 + \rho^2 - 2\rho(x_{a_1} \sin \mu - y_{a_1} \cos \mu) - r_{a_1}^2 = 0. \quad (20)$$

In this equation,  $x_{a_1}$  and  $y_{a_1}$  are coordinates for the tops of flexspline teeth in  $\{x_1 o_1 y_1\}$ . It is only necessary to take the parameter  $u_{a_1}$  which corresponds to the top of the teeth and substitute it into equation (3), and it is then possible to solve. After we have solved for  $x_{a_1}$  and  $y_{a_1}$ , except for  $r_{a_1}$ , the other various quantities are all functions of  $\phi$ . This makes it easy to solve for  $\phi$  from equation (20). This is easier if we let  $\phi$  be equal to  $\phi_{L1}$ . If we again give thought to the matter of gear displacement, as far as the angles of rotation of the axes of symmetry and such other related influences as the thickness of the tops of the gear teeth go, it is possible to solve for the angle  $\phi_{L1}$  which corresponds to the end point of the area of meshing in the first quadrant, that is,

$$\phi_{L1} \approx \varphi_1 + \frac{v}{r_m} + \frac{(r_{a_1} - r_m)\mu + w \operatorname{tg} \alpha_{a_1}}{r_{a_1}} + \frac{S_{a_1}}{2r_{a_1}}, \quad (21)$$

Angle  $\phi_{L1}$  is

$$\phi_{L1} = \varphi_1 + \frac{W_{a_1}}{2r_{a_1}}. \quad (22)$$

In this equation,  $S_{a_1}$  - is the thickness of the top of the teeth on a flexspline

$w_{\alpha 2}$  - is the thickness of the space between the top of the teeth on a rigid gear

From this it is possible to solve for the non-interference condition for the instant when gears are meshing and unmeshing, that is,

$$\phi_{L2} > \phi_{L1}.$$

Due to the fact that  $\mu$  is very small, if we are thinking in terms of making the calculation process easier, then, concerning the original curve which was recommended by this article, when  $\beta=30^\circ$ , then, it is possible to solve for  $H(\varphi)$  on the basis of the equation below, that is,

$$H(\varphi) = \frac{r_{\alpha 2} - r_{\alpha 1}}{w_0}$$

After this, we go to Figure 4 and, from it, obtain a pretty good value for  $\phi_n$ , then, we take  $\varphi_n'$  and substitute it into equation (20), and, by making quite a small number of test calculations, we can, then, solve for the value of  $\phi_n$ .  $H(\varphi)$  is a parameter which is related to radial deformations.

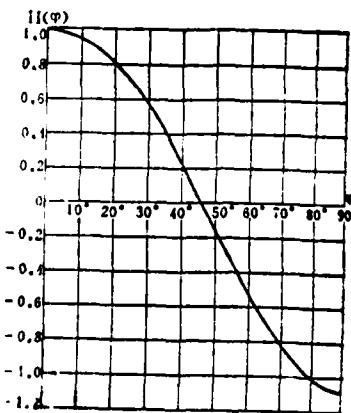


Figure 4

Of course, using computers to do the calculations for the solution of equation (20) is also very convenient. Due to the fact that the function of equation (20) graphs out, in the area of common use, to be very close to a straight line with a very small rate of slope, it is possible to make use of the halving method in the solution.

## 2. Calculations of the Non-Interference Conditions for a Given Meshing Position as Well as for Backlash

If we assume that  $M_1(M_{u1}, Y_{u1})$  are the coordinates for a certain point on the gearing profile of a flexspline within  $\{XOY\}$ , and, if we further assume that we take  $r_m = \sqrt{X_{u1}^2 + Y_{u1}^2}$  to be the radius of arc which intersects with the gearing profile of a neighboring rigid gear, then, we get a point of intersection which is  $M_2(X_{u2}, Y_{u2})$  (Figure 5). In such a case, when the point of meshing which is being considered is in the first quadrant, then, the conditions for non-interference are obviously

$$\begin{cases} X_{u2} - X_{u1} \geq 0, \\ Y_{u2} - Y_{u1} \geq 0. \end{cases} \quad (24)$$

In these inequalities, the values for the coordinates  $M_1, M_2$  are

$$\begin{aligned} X_{u1} &= r_1 \{ \sin[\psi - (u_{u1} - \theta_1)] + u_{u1} \cos \alpha \cos[\psi - (u_{u1} - \theta_1 + \alpha)] \} + \rho \sin \varphi_1 - r_m \sin \psi, \\ Y_{u1} &= r_1 \{ \cos[\psi - (u_{u1} - \theta_1)] - u_{u1} \cos \alpha \sin[\psi - (u_{u1} - \theta_1 + \alpha)] \} + \rho \cos \varphi_1 - r_m \cos \psi, \\ \psi &= \varphi_1 + \mu. \end{aligned} \quad (25)$$

$$\begin{cases} X_{u2} = r_2 \{ \sin[\varphi_2 - (u_{u2} - \theta_2)] + u_{u2} \cos \alpha \cos[\varphi_2 - (u_{u2} - \theta_2 + \alpha)] \}, \\ Y_{u2} = r_2 \{ \cos[\varphi_2 - (u_{u2} - \theta_2)] - u_{u2} \cos \alpha \sin[\varphi_2 - (u_{u2} - \theta_2 + \alpha)] \} \end{cases} \quad (26)$$

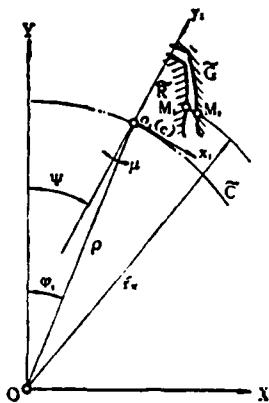


Figure 5

During this solution, first settle on the  $u_{M1}$  which goes with a selected point on the gearing profile of a flexspline. Then, substitute in equation (25), and solve for  $X_{M1}, Y_{M1}$ . From this one can obtain  $r_{M1}$ . Then, from  $r_{M1}$  get a precise value for  $u_{M2}$  on the basis of  $u_{M2} = \arccos \frac{r_2 \cos \alpha}{r_{M1}}$ . Substitute this into equation (26), and get  $X_{M2}, Y_{M2}$ .

The situation in the second quadrant is entirely similar. In such a case, one should take  $(-\varphi)$  and substitute. After this, equation (24) should add a change, but it would be useless to discuss it here.

Because of all this, after one fixes the  $m_1$  which correspond to the coordinates  $M_1$  on the gearing profile of a flexspline, one should then fix a series of values for  $\phi$ . When this is done, then, it is possible to solve for the coordinates  $M_2$  which correspond to different positions of meshing. If one satisfies equation (24), then, there will be no interference. By contrast, if one does not satisfy it, then, interference will occur. If one uses this method, then, it is possible to make test calculations concerning the state of interference at given points of meshing.

When doing these calculations, it is possible to carry them out along the line defined by certain high points on the gear teeth involved. In the general run of situations, it is only necessary to do the calculations for the apical points of the flexspline teeth involved, and that will do.  $\phi$  can be chosen as any value within the range  $0 \sim +90^\circ$ . As a general practice, the unit increment used is  $5^\circ$ , and, depending on the actual situation involved, it is possible to add to the density of the calculations within certain ranges of values of  $\phi$ .

When the results of calculations demonstrate that there will be no occurrence of interference, then, it is an easy thing to figure out the corresponding backlash values. These backlash values can be calculated on the basis of the equation below.

$$j_t \approx \sqrt{(X_{m2} - X_{m1})^2 + (Y_{m1} - Y_{m2})^2}$$

In each place where there is meshing, we calculate the minimum backlash for a certain number of high points on the gear teeth involved, and these values then function as the backlash values for the meshing locations involved.

Concerning the transmission of power, in a situation involving the effects of a moment of bending, the elastic bending deformation of the body of a flexspline will cause the addition of an extra shear displacement to the gear wheel involved. The magnitude of this displacement is determined by the angle of the bending. If one is dealing with a case in which the backlash is excessively small, then, it is possible that this will give rise to interference in the form of elastic distortion. Because of this, as far as the transmission of power is concerned, it is still necessary to take into consideration the amount of backlash reduction,  $j_e$ , which is caused by the elastic deformation of the body of a flexspline. The

value of  $j_e$  can be approximated on the basis of the equation presented below (4), that is,

$$j_e \approx \frac{Mb}{d_i^2 \delta G} \quad (28)$$

In this equation,  $M$  - is the moment of bending  
 $b$  - is the width of the teeth  
 $\delta$  - is the thickness of the walls of a flexspline  
 $G$  - is the modulus of elastic shear

It is necessary to point out that, in the case of harmonic gearing, due to the effects of elastic adjustments and compensations in the meshing of deformed flexsplines, the appearance of a certain amount of interference will most certainly not mean that the transmission will become jammed. However, in the meshing in such a situation, there will be the appearance of excesses, leading to the intensification of losses due to friction, and a reduction in efficiency. Because of this, during the design process, one must do everything possible to avoid it.

### 3. The Control of Backlash

The influence which the magnitude of backlash has on the operation performance of harmonic gear transmissions is relatively large. The backlash requirements associated with other forms of transmission use are also different in some respects. Harmonic transmissions involved in the transmission of very precise movements require that the minimum backlash be zero. In the case of harmonic transmissions used in the transmission of power in servosystems, the requirement will be that backlash will be controlled within the limits of a certain range. On the one hand, if the backlash at the end point of the meshing arc is excessively small, then, one will see the appearance

of elastic deformation interference. On the other hand, if the minimum backlash is excessively large, then, one will see the creation of hopeless return differences, and this will influence the performance of the systems involved in both the static and dynamic modes. Because of these considerations, in the design process, it is necessary to exercise effective control over backlash.

What is referred to as backlash control is the idea that, in a situation in which one is satisfying the requirements for a minimum backlash of a given value, that one makes a reasonable selection of the meshing parameters involved (for example, the coefficients of displacement for flexspline and rigid gear bodies,  $\xi_1$ ,  $\xi_2$ , or the technological section high in the gearing profile involved,  $h_n$ ). Because of the complexity of the functional relationships between the backlash and the meshing parameters in a harmonic transmission, it is difficult to get solutions when one makes use of the orthodox methods of mechanical design as they have been used in the past. It is necessary to assist the process with the addition of a new type of solution method involving an arithmetical mechanical search.

After analysis it is not difficult to see that, actually, the problem of backlash control is this. Given the condition that the meshing parameters involved satisfy a set of constraining conditions, one must contrive to get a value for the difference value ( $j_t - j_c$ ) which is as small as possible. The value  $j_c$  which is used here is a value of backlash which is given on the basis of the operational requirements of the situation involved (the case in which  $j_c = 0$  is also included here). Because of this fact, the solution to this problem can be reduced to the form of a problem in which an optimum value must be achieved within accompanying constraining limits [5]. As far as the transmission of power is concerned, it is also necessary to carry out test calculations at points along the boundary

arc involved in the meshing process.

If we establish the target functions as

$$f(\mathbf{x}) = \begin{cases} |j_i(\mathbf{x}) - j_{i*}|, & \text{when } j_{i*} \neq 0 \\ j_i(\mathbf{x}), & \text{when } j_{i*} = 0 \end{cases} \quad (29)$$

then, in these equations,

$$j_1(\mathbf{x}) = \sqrt{[X_{M2}(x_1, x_3, x_4) - X_{M1}(x_1, x_3, x_4)]^2 + [Y_{M2}(x_1, x_3, x_4) - Y_{M1}(x_1, x_3, x_4)]^2}$$

$\mathbf{x} = \{x_1, x_2, x_3, x_4\}$ , and  $x_1, x_2, x_3, x_4$  respectively represent the chosen values for  $\xi_1, \xi_2, \varphi, h_n$ .

The constraining conditions with  $g_i(\mathbf{x}) \geq 0$  ( $i = 1, 2, 3, \dots, k$ ) are:

(1) There be no production of superimposition interference on the gearing profile. On the basis of equation (24), we obtain

$$g_1(\mathbf{x}): \begin{cases} X_{M2}(x_1, x_3, x_4) - X_{M1}(x_1, x_3, x_4) \geq 0 \\ Y_{M1}(x_1, x_3, x_4) - Y_{M2}(x_1, x_3, x_4) \geq 0 \end{cases}$$

(2) There be no production of excessive curvature interference

$$g_2(\mathbf{x}) = [r_{g2}(x_1) - r_{g1}(x_1)] - (x_1 + w_0) > 0;$$

(3) The height of the technological section of the gearing profile should not exceed its permitted maximum limiting value

$$g_3(\mathbf{x}) = m \left( \frac{\xi_1}{2} + x_1 + 1 \right) - r_{g1}(x_1) - x_1 \geq 0;$$

(4) One must be sure that the depth of meshing not be smaller than  $m$

$$g_4(\mathbf{x}) = x_4 - \frac{1}{2} [r_{s1}(x_2) - r_{s1}(x_1) + m - w_0] > 0,$$

(5) One must be sure that there is a certain definite radial interval between the gear tooth tips of a flexspline and the base of the gear teeth of a rigid gear wheel

$$g_5(\mathbf{x}) = r_{f2}(x_2) - [r_{s1}(x_1, x_4) + w_0 + 0.2m] \geq 0,$$

(6) The tips of gear teeth must not become sharp.

$$g_6(\mathbf{x}): \begin{cases} S_{a1}(x_1, x_4) - 0.25m \geq 0 \\ S_{a2}(x_2, x_4) - 0.25m \geq 0 \end{cases}$$

(7) One must make sure that the wave generator, in the direction of the short axis, can transmit power and smoothly disengage from meshing.

$$g_7(\mathbf{x}) = [r_{s2}(x_2, x_4) + 1.09w_0] - r_{u1}(x_1, x_4) > 0$$

(8) One must insure that the value which is chosen for the backlash is within the range of the meshing arc.

$$g_8(\mathbf{x}) = u_{M2}(x_1, x_3, x_4) - u_{u2}(x_1, x_4) \geq 0$$

In these conditions,  $r_{s1}(x_1), r_{s2}(x_2)$  represent respectively the radii at the start point and end point of the involute curves of a flexspline and a rigid gear. Both of these quantities are functions of their coefficients of displacement [6]. The methods used to express the other various quantities are the same, and their meanings are the same as they were previously.

Because of all this, the problem becomes one of simply being able to find  $x$  inside the working region  $L^* = \{x | g_i(x) \geq 0\} \subset R^n$ . This makes

$$j(\bar{x}) = \min_{x \in L^*} f(x) \quad (30)$$

In this problem,

(1) Because of the fact that the value chosen for  $x$  is only related to  $j_t(x)$  in  $f(x)$ , in order to make the conditional extreme value change to a simple extreme value, one must make

$$j_t(x) = \begin{cases} j_t(x), & x \in L^* \\ N, & x \notin L^* \end{cases}$$

In these equations,  $N$  is a sufficiently large positive number to make  $N \gg j_t(x)$ .

(2) Make use of the initial point,  $x^0$ , of a structure of false random numbers in an even distribution in order to solve for the integral minimum value in this problem.

(3) Make use of the Hooke-Jeeves method for solving for the extreme minimum value of  $f(x)$ .

Concerning the use of this method, as far as the carrying out of the designing of the harmonic transmission of power in certain servosystems goes, the parameters of transmission are

~~$z_1 = 202$~~   $z_1 = 202$ ,  $m = 0.5$ ,  $\alpha = 20^\circ$ . The original curve makes use of the four-force effect form for  $\beta = 30^\circ$  a wall thickness of  $\delta = 0.9$  mm,  $\omega = m = 0.5$ , flexsplines which were finished by the use of hobbing in their machining, and rigid gear wheels which make use of gear cutters in their machining  $z_2 = 60$ ,  $\beta = 0$ . It is required that  $j_c \approx 0$  in order to guard against elastic deformation interference. It is also required that the gear backlash at the end point of the meshing arc be  $j_{cb} = 0.030 \pm 0.005$  mm. After we

make use of electronic computers to optimize our designs, we obtain the following values:  $\xi_1 = 2.801$ ,  $\xi_2 = 2.520$ ,  $h_n = 0.828$  mm,  $r_{a1} = 51.7884$  mm,  $r_{a2} = 50.7234$  mm,  $r_{u1} = 51.6087$  mm,  $r_{u2} = 52.4020$  mm,  $j_{\text{min}} = 6.926 \times 10^{-5}$  mm, and a backlash value at (two characters unreadable) point of  $j_n = 0.03258$  mm. The backlash curve for this is as shown in Figure 6.

To present another example, if a certain piece of electronic equipment precisely determines parameters to transmit power, then, the parameters of it are  $z_1 = 124$ ,  $z_2 = 126$ ,  $m = 0.2$ ,  $\alpha = 20^\circ$ ,  $\omega_0 = m = 0.2$ , the original curve has  $\beta = 30^\circ$  and  $\delta = 0.22$  mm. The flexsplines involved make use of hobbing in their machining, and the rigid gears involved make use of gear cutters in their machining; the gear cutter parameters are  $z_0 = 60$ ,  $\xi_0 = 0.28$ , and it is required that  $j_c = 0$ . The results of these optimized designs are  $\xi_1 = 3.079$ ,  $\xi_2 = 2.506$ ,  $h_n = 0.279$  mm,  $r_{a1} = 13.1457$  mm,  $r_{u1} = 12.7458$  mm,  $r_{a2} = 13.1077$  mm,  $r_{u2} = 13.3875$  mm,  $j_{\text{min}} = 7.57 \times 10^{-5}$  mm.

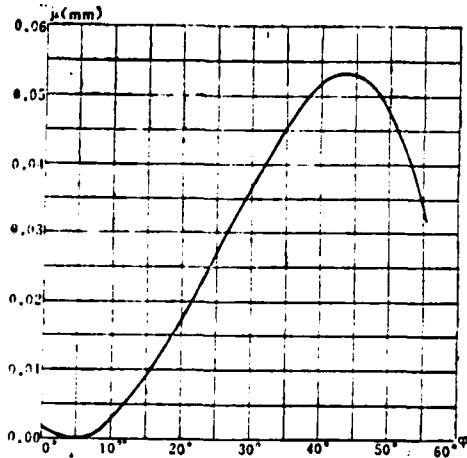


Figure 6

When one makes use of the methods discussed above, it is important to pay attention to the following point, that is, the selection of an increment of search which is reasonable in the

situation at hand. In order to quicken the rate of decrease, it is possible, when the pattern direction is being extended, to add in an extending factor  $T > 1$ ; moreover, on the base of experimental calculations and an appropriate compression, to select a parameter range for initial points. In such a way, accuracy control during calculations can be determined on the basis of actual requirement.

#### IV. A CONSIDERATION OF THE PROBLEM OF DEFORMATIONS OF ORIGINAL CURVES DURING RESEARCH INTO THE TRANSMISSION OF POWER THROUGH HARMONIC TRANSMISSIONS

In the transmission of motive force through transmissions, when one is dealing with the effects of both meshing forces and deforming forces, the original curves of the gear profiles will produce distorted forms and configurations which are described by deviation distance equations (1). In such a situation the loading distribution in the intervals between gear teeth will also give rise to alterations. In order to make precise studies of problems involved with the strengths of power transmitted through harmonic transmissions, it is necessary to find the rules which govern the loading distribution between teeth. At the same time, when doing an analysis of the meshing which goes on in the transmission of power, although it is possible, in situations where one is doing very gross sorts of calculations and/or is involved with elastic deformations, to select the coefficients of displacement, if one needs to get the best possible meshing performance, then, it is necessary to get an accurate solution for the original curves involved, and, with this, to make corrections to the meshing parameters or to correct appropriately the convex gearing profile of the wave generators involved. This section will, while giving simultaneous consideration to the cases of flexsplines and rigid gear wheels as well as to wave generators in terms of their responses to elastic deformation, discuss practical analytical methods for use with the subject of the transmission of power through har-

monic transmissions. In order to make the results of these studies as universally applicable as possible, we will not discuss the situation involving the convex gear forms of wave generators. Other forms of wave generators can all be seen and described as special cases of this generalized situation.

All the experimentation which has been done inside China and abroad concerning this subject demonstrates that the area of meshing involved in the transmission of power, deviations will occur at symmetrical positions toward one side or the other. The configurations of flexsplines will give rise to distortions. There will be sections of the inner walls of these gears where there will be separation between them and the surface of the wave generator involved. In the same type or situation, some sections of wall will adhere to each other very closely. The surfaces of the wave generators concerned will also give rise to distortions. For example, in a double wave transmission, things even go to the point where one sees the appearance of two upper areas of meshing [7], and this causes a new distribution of the loading between gear teeth. In this case, we make use of a type of dispersion set-up [8] in order to make calculations of the configuration of the original curves involved after they are deformed.

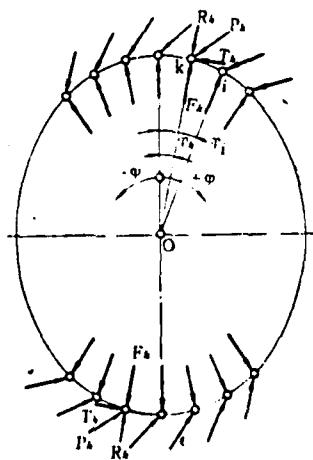


Figure 7

We must first take the loading which is effecting a flexspline and carry out a dispersion within the arc of meshing, taking as the points to be dispersed the nodal points which are the points of intersection between the axes of symmetry of the gears involved and the centerlines of the flexsplines (Figure 7). The pressure which is having an impact on the gears being considered is  $P_k$ . The component forces which correspond to the radial direction and the tangential direction are

$$R_k \approx P_k \sin \alpha_{u_k}, \quad T_k \approx P_k \cos \alpha_{u_k}$$

(One character unreadable) is the angle of the pressure on the gear tips of the flexspline), and the opposing force which the wave generator exerts on the flexspline is  $F_k$ . If we consider the case in which we ignore the effects of the arm of curvature force of the gears involved, then, we take these forces and work with their respective effects on the various individual nodal points ( $k=1,2,\dots,n$ ). If one gives some thought to the total deformation which is impressed on the gears involved under the effects of loading, the radial displacement of the original theoretical curve for the same gear ought to equal the amount of deformation of the wave generator at the same spots. In the same way, the amount of deformation of gears ought to be equal to the amount of meshing excess in the gears. Because of these relationships, it is possible to make the following equations

$$\begin{cases} w_{0k} - w_k = \lambda_{hk} F_k \\ j_{hk} = -\lambda_{hk} T_k \end{cases} \quad (31)$$

Moreover, the relationship between the forces involved and the moments of force is

$$2r_{u_k} \cos \alpha_{u_k} \sum_k P_k = M \quad (32)$$

Equations (31) and (32) are the basic equations used in solving for the unknown forces  $P_k$  and  $F_k$ . In these equations,  $w_k$  is the theoretical radial displacement of a point  $k$  on the centerline of a flexspline, and this is calculated on the basis of equation (2).  $\omega_k$  is the total radial displacement produced at a certain point in the nodal point forces operating there, and the direction of centerline from the back of the component is positive.  $\lambda_{Hk}, \lambda_{Sk}$  are respectively the flexibilities of the wave generators and the corresponding gears involved in the situation. The values of these quantities ought to be verified by experimentation. However, from [9], one can generally recognize the fact that the various points are equal in these respects.

In order to make the calculations more convenient, the backlash  $j_{tk}$  can be represented as the linear forms of  $w_k, v_k, \mu_k$  [10], and, in the situations under study, that means that

$$j_{tk} = j_0 - \frac{(z_2 - z_1)r_{a1} \cdot \phi_k}{z_2} - (r_{a1} - r_m)\mu_k - w_k \operatorname{tg} \alpha_m - v_k. \quad (33)$$

In this equation, the first quantity represents the flexspline before it has been deformed and its initial backlash which is determined by the initially corresponding positions on the axis of symmetry of the flexspline and the axis of symmetry of the rigid gears involved (that is to say, the backlash when  $\phi=0$ ). The second quantity represents the amount of change in the backlash as a response to a change in the angle  $\phi$ . Finally, the third quantity represents the amount of change in the backlash caused by changes in position in radial and shear directions as well as due to rotation of the normal line involved.

Because of the fact that the ratio between the amount of deformation of the flexspline involved is very small compared

to the corresponding radii, for the purposes of our discussion, we will recognize the fact that the principle of superposition is still in effect. On the basis of what is explained in [2 and 11], by using energy principles, we can take  $R_k$ ,  $T_k$  and use  $P_k$  to make the respective substitutions. After one does this, then, it is possible to obtain the equations which represent  $R_k$ ,  $T_k$  at the  $i$ th nodal point under the effects of all the forces which are felt at nodal points, that is to say,  $\omega_1$ ,  $v$  and  $\mu$ .

$$\left\{ \begin{array}{l} \omega_1 = \frac{r_m^3}{\pi E J} \left[ \sum_k F_k \sum_{n=2,4,6,\dots} \frac{\cos n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} - \sum_k P_k \sin \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{\cos n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} \right. \\ \quad \left. - \sum_k P_k \cos \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{\sin n(\varphi_i - \varphi_k)}{n(n^2 - 1)^2} \right] \\ \quad = \frac{r_m^3}{\pi E J} \left[ \sum_k F_k a_{ik} - \sum_k P_k (a_{ik} \sin \alpha_{n1} + b_{ik} \cos \alpha_{n1}) \right], \\ v_1 = - \frac{r_m^3}{\pi E J} \left[ \sum_k F_k \sum_{n=2,4,6,\dots} \frac{\sin n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} - \sum_k P_k \sin \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{\sin n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} \right. \\ \quad \left. + \sum_k P_k \cos \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{\cos n(\varphi_i - \varphi_k)}{n^2(n^2 - 1)^2} \right] \\ \quad = - \frac{r_m^3}{\pi E J} \left[ \sum_k F_k c_{ik} - \sum_k P_k (c_{ik} \sin \alpha_{n1} - d_{ik} \cos \alpha_{n1}) \right], \\ \mu_1 = \frac{r_m^2}{\pi E J} \left[ \sum_k F_k \sum_{n=2,4,6,\dots} \frac{n \sin n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} - \sum_k P_k \sin \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{n \sin n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} \right. \\ \quad \left. + \sum_k P_k \cos \alpha_{n1} \sum_{n=2,4,6,\dots} \frac{\cos n(\varphi_i - \varphi_k)}{(n^2 - 1)^2} \right] \\ \quad = \frac{r_m^2}{\pi E J} \left[ \sum_k F_k e_{ik} - \sum_k P_k (e_{ik} \sin \alpha_{n1} - f_{ik} \cos \alpha_{n1}) \right]. \end{array} \right.$$

(34)

Careful attention should be paid to the fact that, during calculations, the angular starting points are congruent with the long axis of the wave generator. If one takes equation (33) and equation (34) and substitutes them into equation (31), then, after an appropriate exchange of subscripts, it becomes easily possible to get the set of equations for all the nodal points set out below. If we use a matrix form to express this,

we get,

$$\begin{cases} ([A_{ik}] + \lambda_{ik}[E])\{F_k\} - [B_{ik}]\{P_k\} = \{w_{ik}\}, \\ [D_{ik}]\{F_k\} + ([C_{ik}] + \lambda_{ik}\cos\alpha_{ik}[E])\{P_k\} = \{G_k\} \end{cases} \quad (35)$$

In these equations, the parameter matrices are

$$\begin{aligned} [A_{ik}] &= \frac{r_m^2}{\pi E J} [a_{ik}], \\ [B_{ik}] &= \frac{r_m^2}{\pi E J} (\sin\alpha_{ik}[a_{ik}] + \cos\alpha_{ik}[b_{ik}]), \\ [C_{ik}] &= \frac{r_m^2}{\pi E J} \left\{ \sin\alpha_{ik}(\operatorname{tg}\alpha_{ik}[a_{ik}] - [c_{ik}] - \frac{(r_{ik} - r_m)}{r_m}[e_{ik}]) \right. \\ &\quad \left. + \cos\alpha_{ik}(\operatorname{tg}\alpha_{ik}[b_{ik}] + [d_{ik}] - \frac{(r_{ik} - r_m)}{r_m}[f_{ik}]) \right\}, \\ [D_{ik}] &= \frac{r_m^2}{\pi E J} (-\operatorname{tg}\alpha_{ik}[a_{ik}] + [c_{ik}] - \frac{(r_{ik} - r_m)}{r_m}[e_{ik}]), \\ [E] &= \text{(unit matrix)} \end{aligned}$$

Moreover,

$$G_i = \frac{(z_2 - z_1) r_m}{z_2} \varphi_i - j_i$$

From equation (35) and equation (32), it is easily possible to set out the  $(2n + 1)$  equation. When we solve for this, we ought also to take  $j_o$  and look at it as an unknown. On this basis, if we solve the set of linear equations, then, it is possible to solve for the corresponding values of  $j_o$  as well as for the  $2n$  unknown forces  $P_1, P_2, \dots, P_n$  and  $F_1, F_2, \dots, F_n$  which are applied at the  $n$  nodal points. If, after we solve for these quantities, there is the appearance of the conditions

$P_k < 0$  and  $F_k < 0$ , then, this demonstrates the existence of a condition in which suitable contact is not being made. In such a condition, one ought to respectively take the following measures. He should take measures to get rid of the extreme values among all the forces for which  $P_k < 0$  and  $F_k < 0$  are the largest. Then, one should make repeated calculations, and solve for the technological points at which the contact conditions are satisfied. At these points, the following conditions

should apply  $P_k \geq 0$ ,  $F_k \geq 0$ ,  $j_{ik} \leq 0$ ,  $w_{ok} - w_k \geq 0$ . When beginning the calculations, one can determine the individual nodal points on the basis of the central angles which correspond to the arc of meshing (see the previous section for the methods to be used in these calculations). After this, it will be possible to precisely determine the actual technological points on the basis of the solution results. Moreover, it will also be possible to obtain the graph for the loading distribution between the gear teeth involved. This is the foundation for relatively more advanced strength calculations regarding the study of harmonic transmissions.

This method can also be used in the study of such problems as those listed below, that is.

1. The simulation of the loading of harmonic transmissions. First, on the basis of  $j_{ik}$  make a precise determination of the initial corresponding positions of  $\phi_0$  for flexsplines and rigid gears. Concerning rigid gear transmission, after this initial step, take an appropriate increment of length  $\Delta\phi$ , and gradually rotate the flexsplines involved. From this, make a precise determination of  $j_0$ . It is also possible to make direct changes in  $j_0$ . After this, solve for  $P_k$  and  $F_k$  on the basis of equation (35), and, after that, equation (32) acts as the test condition. The process of calculation continues until a force is solved for which satisfies equation (32), then, it stops. In this way, it is possible to see the process of loading in the transmission being studied. Moreover, after all of this is done, it is then possible to obtain the graph for the loading distribution between gear teeth. Of course, during the time of the calculations, all the technological points must, in the same way, satisfy the conditions stated above.

2. Adjusting the convex gearing profile configuration of wave generators. Take the unknown forces which we have already

solved for and substitute them into equation (34). In this way, one can get the configuration of the deformation after the loading of a flexspline. After doing a comparison with the theoretical profile, one should then, on the basis of requirements, carry out the appropriate adjustments in the convex gearing profile of the generator involved. This should make it easier, after the application of loading, to make the configuration of the deformation involved more closely approach the theoretical configuration, and, thereby, improve the meshing performance. However, this operation is relatively complicated, and requires the consideration of many types of factors.

3. On the basis of the values for  $w$ ,  $v$ ,  $\mu$ , which were already obtained, one can apply the methods of the previous section to good advantage or use equation (33) to solve for  $j_n$ . When this is done, then, it is possible to offer a basis for the making of reasonable selections of the backlash and meshing parameters for the transmission of power.

## V. CONCLUSIONS

From the analysis presented above, it can be seen that:

1. This article does not just present a method for the numerical analysis of problems associated with research into the areas of the theoretical meshing of harmonic transmissions, and make use of these methods to rigorously demonstrate the reasonableness of the utilization in harmonic transmissions of involute gearing for which  $\alpha = 20^\circ$ . When one is using this type of gear form, it is only necessary to carry out the appropriate displacements, and that is that. This gives a theoretical basis to the widened use of involute harmonic transmissions. Besides this, the principles which were discussed in this article in the context of the utilization of harmonic gear transmissions with certain original curves, can also be applied to the cases of

harmonic transmissions with gear forms of other types, and this can be done with equal applicability.

2. Concerning the backlash control presented in this article, for the purposes of engineering design, this article presents a type of practical method for selecting the optimum meshing parameters when deciding on a specific value of backlash. This type of method is not only capable of being used in order to control minimum backlash. It is also capable of being used in the control of backlash on the boundaries of the meshing arc so as to guarantee that there will be no occurrence of elastic deformation interference. Theoretical analysis and actual design applications demonstrate that using the meshing parameters which this method precisely quantifies is very satisfactory. Moreover, this method is just as effective when it is used in the selection of meshing parameters for gear-type transmissions and in the design of complex-type transmissions.

3. Concerning how this article gave overall consideration to the ways in which elastic distortion occurs in the components of transmissions, it discussed several important problems relating to research in the area of harmonic transmissions, for example, the loading distribution between gears, loading simulation, the adjustment of the convex gearing profiles of wave generators, the selection of backlash, and other problems of a similar type. Actual calculations demonstrate that it is convenient to make use of conjugate slope methods of solution when working with these methods. However, this type of operation still needs to wait on a large amount of experimental work before it can serve as a replacement for current methods.

Finally, it is still necessary to point out the following. In the process of this research, the large amount of calculation work involved met with the ardent help of such comrades, from the National Yellow River Machine Plant, as Cheng Bao-yuan, Fan Shou-liang, Li Fu-ying and others. For this, the author wishes to express his gratitude.

## References

- [1] Musser, C. W., The Harmonic Drive, Breakthrough in Mechanical Drive Design, *Machine Design*, Vol. 32, No. 8, 1960, pp. 160-173.
- [2] Timoshenko, S., Woinowsky-Kreiger, S., *Theory of Plates and Shells*, McGraw-Hill Book Company, Inc. (1959), pp. 501-507.
- [3] Колчина, Н. И., *Аналитический Расчет Плоских и Пространственных Зашеплений*, Машгиз (1959), с. 10-18.
- [4] Шувалов, С. А., *Расчет Волновых Передач с Учетом Податливости*, Звеньев, *Вестник Машиностроения*, № 6, 1974, с. 46-51.
- [5] Mangasarian, O. L., *Techniques of Optimization*, AD 727205 (1971).
- [6] Гавриленко, В. А., *Зубчатые Передачи в Машиностроении*, Машгиз (1962) с. 175-181, 284-286.
- [7] Волков, Д. П., Крайнев, А. Ф., *Волновые Зубчатые Передачи*, Изд. Техника (1976), с. 102-111.
- [8] Ковалев, Н. А., *Некоторые Вопросы Теории Волновых Зубчатых Передач*, *Машиноведение*, № 2, 1973, с. 48-55.
- [9] Гиазбург, Е. Г., *Исследование Коэффициентов Упругих Перемещений Зубьев Гибкого Колеса, Зубчатые и Червячные Передачи*, *Машиностроение* (1974), с. 220-222.
- [10] Ковалев, Н. А., *О Распределении Нагрузки по Зубьям в Волновой Зубчатой Передаче*, *Машиноведение*, № 5, 1974, с. 44-49.
- [11] Филиппенков, А. Л., *Исследование Деформированного и Напряженного Состояния Зубчатых Колес Планетарных Передач. Зубчатые и Червячные Передачи* (1974), с. 159-171.

## Summary

# A Method of Evaluation of the Torsional Rigidity and the Third Stress Intensity Factor of Prismatical Bar of Rectangular Cross-section with Cracks

Chen Yizhou

In this paper, a method of evaluation of the torsional rigidity and the third stress intensity factor of rectangular bar with cracks is discussed. In all the three cases considered here, the crack is perpendicular to the edge of the rectangular section of the prismatical bar. The three cases are: (1) one straight-line crack not originating from the midpoint of one edge, (2) one straight-line crack originating from the midpoint of one edge, (3) two straight-line cracks of equal length originating from the midpoints of opposite edges.

Case (2) was solved by Westmann<sup>[1]</sup>. In his paper, he reduced the problem to a mixed boundary-value problem for Poisson's equation in two dimensions. Utilization of Sneddon's double-series method of solution then permits the further reduction of the problem to the numerical solution of a Fredholm integral equation of the second kind. We can see that, the method used in Westmann's paper is rather complicated not only in the theoretical analysis but also in the computation work.

Sih pointed out that<sup>[2]</sup>, if the conformal mapping technique is used, in the case of complex geometries, the determination of the entire stress field of cross-section with cracks might require a great effort.

For the purpose of introducing something better than the existing methods, the author proposes a new method, which is based on both the compliance method in fracture mechanics and the continuation theorem of harmonic functions. As can be seen later, the method in this paper is very effective, provided the crack is perpendicular to the edge of the rectangular section.

It is well known that, in linear elastic fracture mechanics, the inverse of the torsional rigidity is the compliance, and by differentiating the compliance with respect to crack length,  $K_3$  can be obtained. This is the so-called compliance method.

As we know, the torsion problem in elasticity can be turned into a

Dirichlet's problem of Laplace's equation. The rectangular cross-section with cracks is divided into several rectangles, and along the dividing lines undetermined functions are placed. Now the Dirichlet's problems for all the rectangles can be solved. The undetermined functions can be determined by requiring the normal derivatives of two harmonic functions to be equal to each other along the dividing lines. The solution so obtained satisfies the given boundary value problem. The condition of continuation of the harmonic function always leads to the problem of solving a linear algebraic system. After the Dirichlet's problem is solved, the calculation of the torsional rigidity and  $K_3$  will be easy and convenient.

Corresponding to these three cases, the numerical results for torsional rigidity and  $K_3$  are listed in tables 1, 2, 3, 4, 5 and 6 respectively.

A Method of Evaluation of the Torsional Rigidity  
and the Third Stress Intensity Factor of a Prismatic  
Bar of Rectangular Cross-section with Cracks

/161

Chen Yizhou

In this paper a method of calculation of the torsional rigidity and the third stress intensity factor of a rectangular bar with cracks is discussed. Under the three conditions discussed in this paper, the cracks are perpendicular to the edges of the rectangular cross-section of the prismatic bar. The three cases covered in this paper are: 1) a single crack not originating from the center of the edges of the rectangular cross-section; 2) a single crack originating from the midpoint of one edge of the rectangular cross-section, and 3) two cracks of equal length originating from the midpoints of opposite edges of the rectangular cross-section.

The problem was solved by Westmann under condition 2), [1]. In his paper, he reduced this problem to a two-dimensional Poisson equation with mixed boundary values. This problem is used to reduce the problem further to a Fredholm integral equation of the second kind. It is obvious, not only in terms of theoretical analysis but also based on the calculation process, that the method used in Westmann's paper is very complicated.

On the other hand, Sih [2] pointed out that the use of the conformal mapping technique to determine the entire stress field of the cross-section with cracks requires a tremendous amount of effort in the case of complex geometrics.

In order to improve the existing methods in solving the torsion problem of prismatic bars with cracks, we are going to propose a new technique which is based on the compliance method in fracture mechanics and the continuation theorem of harmonic functions. It can be demonstrated later that our method is very effective as long as the crack is perpendicular to the edge of the rectangular cross-section.

It is well known that the torsion problem in elastic mechanics can be turned into a Dirichlet problem of a Laplace equation. The cracked rectangular cross-section is divided into several rectangular regions. Undetermined functions are also placed along the dividing lines. The Dirichlet problems can be solved for all the rectangles involved. The undetermined functions can be determined by letting the normal derivatives of the two harmonic functions be equal to each other along the dividing lines. This method provides real solutions satisfying the given boundary value conditions. Harmonic functions and continuation conditions usually require the solving of a series of linear algebraic equations. Once the Dirichlet problem is solved, the calculation of torsional rigidity and  $K_3$  becomes a direct and simple task.

Corresponding to the three conditions discussed above, the numerical values of the torsional rigidity and  $K_3$  are expressed in Tables 1, 2, 3, 4, 5 and 6

I. TORSIONAL RIGIDITY AND  $K_3$  STRESS INTENSITIY FACTOR OF CRACKS NOT PASSING THROUGH THE MIDPOINT OF THE EDGE OF THE RECTANGULAR CROSS-SECTION.

The rectangular cross-section of a bar with a crack of one edge is discussed in this section (Figure 1). A method to calculate the torsional rigidity D and stress intensity

factor  $K_3$  is considered.

It is well known that in the torsion problem of an elastic cylinder, the stress function  $\phi(x, y)$  obeys the following [3,4]:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = -2, \quad (1)$$

$$\phi|_{\gamma} = 0.$$

where  $L$  represents the edges of the cross-section. The torsional rigidity is:

$$D = \mu J, \quad J = 2 \iint_R \phi(x, y) dx dy. \quad (2)$$

where  $\mu$  is the shear elastic modulus.

Introducing a new function  $u(x,y)$ , we get

$$\phi(x, y) = -x^2 + u(x, y), \quad (3)$$

Substituting (3) into (1) we get harmonic functions

$$u(x, y)$$

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0, \quad (4)$$

$$u|_L = x^2.$$

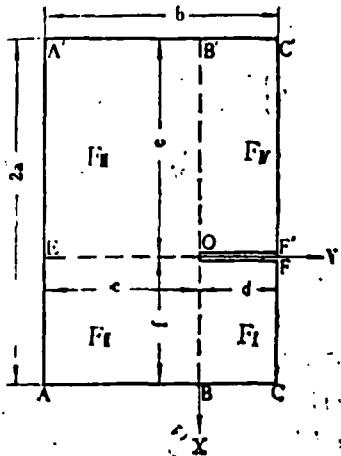


Figure 1. The Geometrical Shape of the Rectangular Cross-section with crack.

In the cross-section shown in Figure 1, it is possible to divide into four rectangles along  $OB, OE, OB'$  directions. These regions are designated as  $F_x, F_{\bar{x}}, F_{\bar{\bar{x}}}$  and  $F_{\bar{\bar{\bar{x}}}}$ , with their corresponding stress functions  $\phi(x, y)$  and  $u(x, y)$  as  $\phi_x, \phi_{\bar{x}}, \phi_{\bar{\bar{x}}}, \phi_{\bar{\bar{\bar{x}}}}$  and  $u_x, u_{\bar{x}}, u_{\bar{\bar{x}}}, u_{\bar{\bar{\bar{x}}}}$ , respectively.

/162

Let us set:  $u(x, y)|_{x=0} = g(y), \quad (-c \leq y \leq 0)$   
 $u(x, y)|_{y=0} = -ex + h(x), \quad (-e \leq x \leq 0)$   
 $u(x, y)|_{y=0} = fx + f(x), \quad (0 \leq x \leq f) \quad (5)$   
 $W(x) = x^2 - fx, \quad V(x) = x^2 + ex.$

where  $f(x), h(x)$  and  $g(x)$  are undetermined functions.

It is not difficult to see that the fixed solution problems of harmonic functions  $u_x, u_{\bar{x}}, u_{\bar{\bar{x}}}$  and  $u_{\bar{\bar{\bar{x}}}}$  in regions  $OBCE, OEAB, OB'A'E$  and  $OF'C'B'$  are:

$$\begin{aligned} \frac{\partial^2 u_x}{\partial x^2} + \frac{\partial^2 u_x}{\partial y^2} &= 0, & (0 < x < f, 0 < y < d) \\ u_x(x, d) &= fx + W(x), & (0 \leq x \leq f) \\ u_x(f, y) &= f^2, & (0 \leq y \leq d) \\ u_x(x, 0) &= fx + f(x), & (0 \leq x \leq f) \\ u_x(0, y) &= 0. & (0 \leq y \leq d) \end{aligned} \quad (6)$$

and

$$\begin{aligned} \frac{\partial^2 u_{\bar{x}}}{\partial x^2} + \frac{\partial^2 u_{\bar{x}}}{\partial y^2} &= 0, & (0 < x < f, -c \leq y \leq 0) \\ u_{\bar{x}}(x, 0) &= fx + f(x), & (0 \leq x \leq f) \\ u_{\bar{x}}(f, y) &= f^2, & (-c \leq y \leq 0) \\ u_{\bar{x}}(x, -c) &= fx + W(x), & (0 \leq x \leq f) \\ u_{\bar{x}}(0, y) &= g(y). & (-c \leq y \leq 0) \end{aligned} \quad (7)$$

and

$$\begin{aligned} \frac{\partial^2 u_{\bar{\bar{x}}}}{\partial x^2} + \frac{\partial^2 u_{\bar{\bar{x}}}}{\partial y^2} &= 0, & (-e < x < 0, -c \leq y \leq 0) \\ u_{\bar{\bar{x}}}(x, 0) &= -ex + h(x), & (-e \leq x \leq 0) \\ u_{\bar{\bar{x}}}(0, y) &= g(y), & (-c \leq y \leq 0) \\ u_{\bar{\bar{x}}}(x, -c) &= -ex + V(x), & (-e \leq x \leq 0) \\ u_{\bar{\bar{x}}}(-e, y) &= e^2, & (-c \leq y \leq 0) \end{aligned} \quad (8)$$

and

$$\begin{aligned}
 \frac{\partial^2 u_n}{\partial x^2} + \frac{\partial^2 u_n}{\partial y^2} &= 0, & (-e < x < 0, 0 < y < d) \\
 u_n(x, d) &= -ex + V(x), & (-e \leq x \leq 0) \\
 u_n(0, y) &= 0, & (0 \leq y \leq d) \\
 u_n(x, 0) &= -ex + h(x), & (-e \leq x \leq 0) \\
 u_n(-e, y) &= e^y. & (0 \leq y \leq d)
 \end{aligned} \tag{9}$$

It is also not difficult to see that solutions exist for the above four equations:

$$\begin{aligned}
 u_1(x, y) &= fx + \sum_{n=1}^{\infty} f_n \frac{sh(n\pi(d-y)/f)}{sh(n\pi d/f)} \sin(n\pi x/f) + \\
 &+ \sum_{n=1}^{\infty} W_n \frac{sh(n\pi y/f)}{sh(n\pi d/f)} \sin(n\pi x/f). \quad (0 \leq x \leq f, 0 \leq y \leq d) \tag{10}
 \end{aligned}$$

$$\begin{aligned}
 u_2(x, y) &= fx + \sum_{n=1}^{\infty} g_n \frac{sh(n\pi(f-x)/c)}{sh(n\pi f/c)} \sin(n\pi(y+c)/c) + \\
 &+ \sum_{n=1}^{\infty} W_n \frac{sh(-n\pi y/f)}{sh(n\pi c/f)} \sin(n\pi x/f) + \\
 &+ \sum_{n=1}^{\infty} f_n \frac{sh(n\pi(y+c)/f)}{sh(n\pi c/f)} \sin(n\pi x/f). \quad (0 \leq x \leq f, -c \leq y \leq 0) \tag{11}
 \end{aligned}$$

$$\begin{aligned}
 u_3(x, y) &= -ex + \sum_{n=1}^{\infty} V_n \frac{sh(-n\pi y/e)}{sh(n\pi c/e)} \sin(n\pi(x+e)/e) + \\
 &+ \sum_{n=1}^{\infty} h_n \frac{sh(n\pi(y+c)/e)}{sh(n\pi c/e)} \sin(n\pi(x+e)/e) + \\
 &+ \sum_{n=1}^{\infty} g_n \frac{sh(n\pi(x+e)/c)}{sh(n\pi e/c)} \sin(n\pi(y+c)/c). \quad (-e \leq x \leq 0, -c \leq y \leq 0) \tag{12} \quad /163
 \end{aligned}$$

$$\begin{aligned}
 u_4(x, y) &= -ex + \sum_{n=1}^{\infty} h_n \frac{sh(n\pi(d-y)/e)}{sh(n\pi d/e)} \sin(n\pi(x+e)/e) + \\
 &+ \sum_{n=1}^{\infty} V_n \frac{sh(n\pi y/e)}{sh(n\pi d/e)} \sin(n\pi(x+e)/e). \quad (-e \leq x \leq 0, 0 \leq y \leq d) \tag{13}
 \end{aligned}$$

In the above equations,  $f_n, g_n, h_n, W_n, V_n$  and  $\Delta_n$  are the following:

/163

$$\begin{aligned} f_n &= \frac{2}{f} \int_0^f f(x) \sin(n\pi x/f) dx, \quad g_n = \frac{2}{c} \int_{-c}^0 g(y) \sin(n\pi(y+c)/c) dy, \\ h_n &= \frac{2}{e} \int_{-e}^0 h(x) \sin(n\pi(x+e)/e) dx, \\ W_n &= -\frac{8f^2}{n^3\pi^3} \Delta_n, \quad V_n = -\frac{8e^2}{n^3\pi^3} \Delta_n, \quad \Delta_n = \frac{1+(-1)^{n+1}}{2}, \quad n=1,2,\dots \end{aligned} \quad (14) \quad /164$$

According to the continuation theorem of harmonic functions, [3] along the sections OE, OE' and OB' there should be:

$$\frac{\partial u_x}{\partial y} \Big|_{y=0} = \frac{\partial u_x}{\partial y} \Big|_{y=0}. \quad (0 \leq x \leq f) \quad (15)$$

$$\frac{\partial u_x}{\partial x} \Big|_{x=0} = \frac{\partial u_x}{\partial x} \Big|_{x=0}. \quad (-c \leq y \leq 0) \quad (16)$$

$$\frac{\partial u_x}{\partial y} \Big|_{y=0} = \frac{\partial u_x}{\partial y} \Big|_{y=0}. \quad (-e \leq x \leq 0) \quad (17)$$

Now let

$$F_m = -mf_n, \quad G_m = (-1)^m mg_n, \quad H_m = (-1)^m mh_n. \quad m=1,2,\dots \quad (18)$$

If we substitute equations (10), (11), (12), and (13) into (15), (16), and (17) and then execute Fourier series expansion on both sides, the algebraic equations corresponding to  $F_m$ ,  $G_m$ , and  $H_m$  can be obtained by also taking (18) into consideration.

$$F_m = I_m + O_m \sum_{n=1}^{\infty} \frac{G_n}{(nf)^2 + (mc)^2}, \quad m=1,2,3,\dots \quad (19)$$

$$G_m = J_m + P_m \sum_{n=1}^{\infty} \frac{F_n}{(nc)^2 + (mf)^2} + L_m \sum_{n=1}^{\infty} \frac{H_n}{(nc)^2 + (me)^2}, \quad m=1,2,3,\dots \quad (20)$$

$$H_m = K_m + Q_m \sum_{n=1}^{\infty} \frac{G_n}{(ne)^2 + (mc)^2}, \quad m = 1, 2, 3, \dots \quad (21)$$

where ( $I_m, O_m, J_m, P_m, L_m, K_m$  and  $Q_m$ ) are the following known coefficients:

/164

$$\begin{aligned} I_m &= \frac{8f^3}{m^2\pi^3} \frac{EXP(-m\pi c/f) + EXP(-m\pi d/f)}{1 + EXP(-m\pi b/f)} \Delta_m, \\ O_m &= \frac{2mcf}{\pi} \frac{1}{cth(m\pi c/f) + cth(m\pi d/f)}, \\ J_m &= \left[ -\frac{2(e+f)c}{m\pi^2} + (-1)^m \frac{4c^2}{\pi^6 m^2} (th(m\pi f/2c) + th(m\pi e/2c)) \right] \cdot \\ &\quad \cdot \frac{1}{cth(m\pi f/c) + cth(m\pi e/c)}, \\ P_m &= \frac{2mcf}{\pi} \frac{1}{cth(m\pi f/c) + cth(m\pi e/c)}, \\ L_m &= \frac{2mce}{\pi} \frac{1}{cth(m\pi f/c) + cth(m\pi e/c)}, \\ K_m &= \frac{8e^3}{m^2\pi^3} \frac{EXP(-m\pi c/e) + EXP(-m\pi d/e)}{1 + EXP(-m\pi b/e)} \Delta_m, \\ Q_m &= \frac{2mce}{\pi} \frac{1}{cth(m\pi c/e) + cth(m\pi d/e)}. \quad m = 1, 2, \dots \end{aligned} \quad (22)$$

(22)

If we substitute equations (10), (11), (12) and (13) into (3), and then substitute (3) into (2) and carry out a fixed integration calculation followed by turning  $f_m, g_m$  and  $h_m$  into  $F_m, G_m$ , and  $H_m$  using equation (18), the torsional rigidity can finally be obtained.

$$\begin{aligned} D &= \mu J, \\ J &= \frac{b}{3} (e^3 + f^3) - \frac{4c^2}{\pi^2} \sum_{n=1,3}^{\infty} \frac{G_n}{n^3} \left( th \frac{n\pi e}{2c} + th \frac{n\pi f}{2c} \right) - \\ &\quad - \frac{4f^2}{\pi^2} \sum_{n=1,3}^{\infty} \frac{F_n}{n^3} \left( th \frac{n\pi d}{2f} + th \frac{n\pi c}{2f} \right) - \frac{32f^4}{\pi^6} \sum_{n=1,3}^{\infty} \frac{1}{n^6} \left( th \frac{n\pi c}{2f} + th \frac{n\pi d}{2f} \right) - \\ &\quad - \frac{4e^2}{\pi^2} \sum_{n=1,3}^{\infty} \frac{H_n}{n^3} \left( th \frac{n\pi d}{2e} + th \frac{n\pi c}{2e} \right) - \frac{32e^4}{\pi^6} \sum_{n=1,3}^{\infty} \frac{1}{n^6} \left( th \frac{n\pi c}{2e} + th \frac{n\pi d}{2e} \right). \end{aligned} \quad (23)$$

Summarizing our discussion, we can obtain the torsional rigidity by solving  $F_m$ ,  $G_m$  and  $H_m$  using equations (19), (20), and (20) and then plugging into the above equation.

From common theory in fracture mechanics we know that the following equation is valid:

$$G_s = \frac{M^2}{2} \frac{\delta}{\delta(d)} \left( \frac{1}{\mu J(d)} \right) = \frac{K_3^2}{2\mu} , \quad (24)$$

where  $G_s$  is the energy release rate,  $M$  is the torque on the end of the bar,  $1/\mu J(d)$  is the softness of the cracked bar,  $d$  is the length of the crack and  $\delta$  is the symbol for differentiation. From the above equation, we get

$$K_3 = M \sqrt{\frac{\delta}{\delta(d)} \left( \frac{1}{J(d)} \right)} . \quad (25)$$

The solutions to equations (19), (20) and (21) are obtained using Seidel's iteration technique with a 15 term cutoff. The corresponding  $F_m$ ,  $G_m$ , and  $H_m$  are then substituted into (23) to obtain the torsional rigidity  $D$  as shown in Table 1. From (25) using a three point differential method the values of  $K_3$  are obtained as presented in Table 2 (Note: the " $K_3$ " used in this paper is  $\sqrt{\pi}$  times greater than those in other books). For Tables 3 and 4,  $b = 2a$ ; which means that the calculations were made for square cross-section with cracks. Using ALGOL-60 language and a DJS-21 electronic computer, the values reported in this paper were determined.

### II. TORSIONAL RIGIDITY AND $K_3$ STRESS INTENSITY FACTOR OF CRACKS ORIGINATING FROM THE MIDPOINTS OF THE EDGES OF A RECTANGULAR CROSS-SECTION

When a crack originates from the midpoint of an edge of the rectangle (Figure 2), it corresponds to a special case in

the previous section (i.e.,  $a = e = f$ ). From symmetry, we find that  $f(x) = h(-x)$  ( $0 \leq x \leq a$ ). Based on equations (14) and (18), we get

$$f_m = (-1)^{m-1} h_m, \quad F_m = H_m, \quad m = 1, 2, \dots$$

In this case, equations (19), (20), and (21) can be simplified as:

$$F_m = I_m + O_m \sum_{n=1}^{\infty} \frac{G_n}{(na)^2 + (mc)^2}, \quad m = 1, 2, \dots \quad (26)$$

$$G_m = J_m + 2P_m \sum_{n=1}^{\infty} \frac{F_n}{(nc)^2 + (ma)^2}, \quad m = 1, 2, \dots \quad (27)$$

where  $I_m$ ,  $O_m$ ,  $J_m$ , and  $P_m$  were obtained from (22) and they are:

$$\begin{aligned} I_m &= \frac{8a^2}{\pi^2 m^2} \frac{EXP(-m\pi c/a) + EXP(-m\pi d/a)}{1 + EXP(-m\pi b/a)}, \\ O_m &= \frac{2mac}{\pi} \frac{1}{cth(m\pi c/a) + cth(m\pi d/a)}, \\ J_m &= \left( -a + \frac{(-1)^m 2c}{\pi m} \operatorname{th} \frac{m\pi a}{2c} \right) \cdot \frac{2c}{\pi^2 m} \operatorname{th} \frac{m\pi a}{c}, \\ P_m &= \frac{mac}{\pi} \operatorname{th} \frac{m\pi a}{c}. \quad m = 1, 2, \dots \end{aligned}$$

(28)

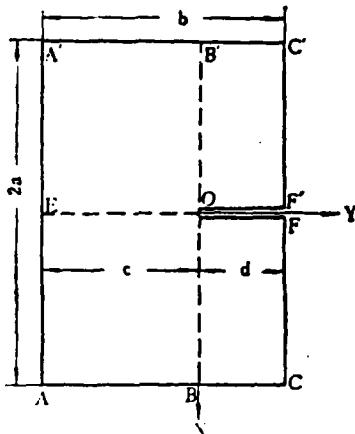


Figure 2.

The torsional rigidity  $D$  can be obtained by simplifying equation (23) as below:

$$D = \mu J,$$

$$J = \frac{2a^3b}{3} - \frac{8a^2}{\pi^2} \sum_{n=1,3}^{\infty} \frac{1}{n^3} \left( \frac{8a^2}{\pi^2 n^2} + F_n \right) \left( th \frac{n\pi c}{2a} + th \frac{n\pi d}{2a} \right) - \frac{8c^2}{\pi^2} \sum_{n=1,3}^{\infty} \frac{G_n}{n^3} th \frac{n\pi a}{2c}. \quad (29)$$

Similar to the previous section, fifteen terms were taken into account when using the Seidel iterative method to solve equations (26) and (27) to obtain the torsional rigidity. The  $K_3$  stress intensity factor can also be obtained using (25). The results are shown in Tables 3 and 4. They were found to be identical to those calculated by Westmann [1] [8]. We also used a boundary matching method to calculate this problem which offered a similar result. The technique used here is actually more superior to the boundary matching method.

III. TORSIONAL RIGIDITY AND  $K_3$  STRESS INTENSITY FACTOR  
FOR SYMMETRIC CRACKS ON OPPOSITE SIDES.

When two cracks originate symmetrically from the midpoints of two opposite edges of the rectangle (Figure 3), equations (1), (2), (3) and (4) still apply. The coordinate system and the dimensions are shown in Figure 3. Just as in the first section, let

$$\begin{aligned} u(x, y) \Big|_{x=0} &= g(y), \quad (-c \leq y \leq c) \\ u(x, y) \Big|_{y=c} &= ax + f(x), \quad (0 \leq x \leq a) \\ W(x) &= x^2 - ax \quad . \end{aligned} \quad (30)$$

where  $f(x)$  and  $g(y)$  are undetermined functions. Due to the symmetric nature of the problem  $g(y)$  is an even function.

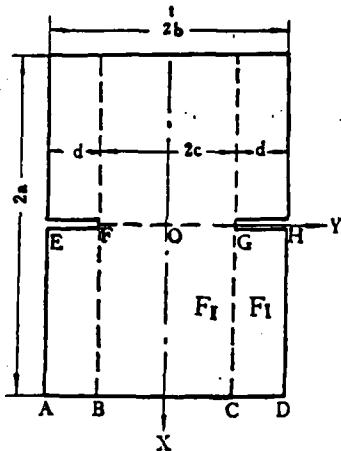


Figure 3

/166

It is apparent that the fixed solutions to the harmonic functions  $u_1(x, y)$  and  $u_2(x, y)$  in regions GHDC and FGCB obey the following:

$$\begin{aligned}
 \frac{\partial^2 u_1}{\partial x^2} + \frac{\partial^2 u_1}{\partial y^2} &= 0, & (0 < x < a, c < y < b) \\
 u_1(x, b) &= ux + W(x), & (0 \leq x \leq a) \\
 u_1(a, y) &= a^2, & (c \leq y \leq b) \\
 u_1(x, c) &= ax + f(x), & (0 \leq x \leq a) \\
 u_1(0, y) &= 0. & (c \leq y \leq b)
 \end{aligned} \tag{31}$$

and

$$\begin{aligned}
 \frac{\partial^2 u_2}{\partial x^2} + \frac{\partial^2 u_2}{\partial y^2} &= 0, & (0 < x < a, -c < y < c) \\
 u_2(x, c) &= u_2(x, -c) = ax + f(x), & (0 \leq x \leq a) \\
 u_2(0, y) &= g(y), & (-c \leq y \leq c) \\
 u_2(a, y) &= a^2. & (-c \leq y \leq c)
 \end{aligned} \tag{32}$$

The solutions to the above two equations are:

$$\begin{aligned}
 u_1(x, y) &= ax + \sum_{n=1}^{\infty} f_n \frac{sh(n\pi(b-y)/a)}{sh(n\pi d/a)} \sin(n\pi x/d) + \\
 &+ \sum_{n=1}^{\infty} W_n \frac{sh(n\pi(y-c)/a)}{sh(n\pi d/a)} \sin(n\pi x/a). \\
 & (0 \leq x \leq a, c \leq y \leq b)
 \end{aligned} \tag{33}$$

$$\begin{aligned}
 u_2(x, y) &= ax + \sum_{n=1,3}^{\infty} g_n \frac{sh(n\pi(a-x)/2c)}{sh(n\pi a/2c)} \sin(n\pi(y+c)/2c) \\
 &+ \sum_{n=1}^{\infty} f_n \frac{ch(n\pi y/a)}{ch(n\pi c/a)} \sin(n\pi x/a). \\
 & (0 \leq x \leq a, -c \leq y \leq c)
 \end{aligned} \tag{34}$$

where  $f_n$ ,  $g_n$ , and  $W_n$  are

$$\begin{aligned}
 f_n &= \frac{2}{a} \int_0^a f(x) \sin(n\pi x/a) dx, \quad g_n = \frac{1}{c} \int_{-c}^c g(y) \sin(n\pi(y+c)/2c) dy, \\
 W_n &= -\frac{8a^2}{n^3 \pi^3} A_n, \quad A_n = (1 + (-1)^{n+1})/2, \quad n = 1, 2, \dots
 \end{aligned} \tag{35}$$

Based on the continuation theorem of the harmonic function, the following conditions exist along sections GC and FG (see Figure 3):

$$\frac{\partial u_x}{\partial y} \Big|_{y=0} = \frac{\partial u_x}{\partial y} \Big|_{y=c}, \quad (0 \leq x \leq a) \quad (36)$$

$$\frac{\partial u_x}{\partial x} \Big|_{x=0} = 0. \quad (-c \leq y \leq c) \quad (37)$$

Now, let us assume that

$$F_m = -m f_m, \quad m = 1, 2, \dots \quad G_m = -m g_m, \quad m = 1, 3, \dots \quad (38)$$

Plugging (33) and (34) into (36) and (37), expanding both sides into Fourier series, and using the above equation, the following equations relating to  $F_m$  and  $G_m$  can be obtained:

$$F_m = H_m + J_m \sum_{n=1,3}^{\infty} \frac{G_n}{(na)^2 + (2mc)^2}, \quad m = 1, 2, \dots \quad (39)$$

$$G_m = I_m + K_m \sum_{n=1}^{\infty} \frac{F_n}{(2nc)^2 + (na)^2}. \quad m = 1, 3, \dots \quad (40)$$

where

$$\begin{aligned} H_m &= \frac{8a^2}{\pi^3 m^2} \frac{ch(m\pi c/a)}{ch(m\pi b/a)} \Delta_m, \quad m = 1, 2, \dots \\ J_m &= \frac{4mca}{\pi} \frac{1}{th(m\pi c/a) + cth(m\pi d/a)}, \quad m = 1, 2, \dots \\ I_m &= -\frac{8ac}{m\pi^2} th(m\pi a/2c), \quad m = 1, 3, \dots \\ K_m &= \frac{8mac}{\pi} th(m\pi a/2c). \quad m = 1, 3, \dots \end{aligned} \quad (41)$$

Substituting (33) and (34) into (3) and then plugging (3) into (2), followed by integration, and turning  $f_m$  and  $g_m$  into  $F_m$  and  $G_m$ , the torsional rigidity  $D$  can finally be obtained as:

$$D = \mu J,$$

$$J = \frac{4b a^4}{3} - \frac{32c^3}{\pi^3} \sum_{n=1,3}^{\infty} \frac{G_n}{n^3} \cdot h \frac{n\pi a}{4c} - \frac{16a^4}{\pi^4} \sum_{n=1,3}^{\infty} \frac{F_n}{n^3} \left( th \frac{n\pi c}{a} + th \frac{n\pi d}{2a} \right) \\ - \frac{128a^4}{\pi^6} \sum_{n=1,3}^{\infty} \frac{1}{n^6} th \frac{n\pi d}{2a} . \quad (42)$$

Similar to the first section, using the Seidel iteration method to solve (39) and (40) and taking 15 terms into account, the torsional rigidity can be obtained. In equation (25), by changing d into 2d the factor  $K_3$  can be determined. The results are shown in Tables 5 and 6.

Table 1. Table of Torsional Rigidity Coefficient  
 $A\left(\frac{d}{b}, \frac{f}{b}\right)$  of a Rectangular Cross-section  
 with Cracks.

/167

$d/b (0, 0.05, \dots, 0.95)$   
 $f/b (0.1, 0.2, 0.3, 0.4, 0.5)$

$d/b$ $f/b$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.1	1405.8	1401.3	1391.7	1378.2	1361.8	1343.1	1322.7	1301.3	1279.4	1257.6
0.2	1405.8	1395.5	1370.0	1342.8	1307.2	1268.1	1227.0	1185.2	1143.8	1103.3
0.3	1405.8	1390.9	1357.7	1312.9	1260.7	1204.4	1146.2	1088.1	1031.3	977.00
0.4	1405.8	1388.0	1348.0	1293.8	1231.1	1163.5	1094.2	1026.3	958.61	895.82
0.5	1405.8	1387.1	1344.9	1287.6	1221.1	1149.6	1076.4	1003.8	933.84	867.86
$d/b$ $f/b$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.1	1236.3	1216.2	1197.6	1180.9	1166.6	1155.0	1146.3	1140.4	1137.2	1136.1
0.2	1065.1	1029.8	998.13	970.63	947.78	929.86	916.89	908.56	904.23	902.74
0.3	928.44	880.50	839.78	805.20	778.96	755.19	739.70	729.92	724.93	723.25
0.4	837.47	785.10	739.27	700.35	669.00	645.03	628.11	617.53	612.18	610.38
0.5	807.11	752.55	704.92	664.74	632.26	607.50	590.07	579.22	573.72	571.89

Equation:  $D = \mu J, \quad J = A\left(\frac{d}{b}, \frac{f}{b}\right) \cdot 10^{-4} \cdot b^4$ .

Example: 若  $\frac{d}{b} = 0.35, \frac{f}{b} = 0.2$ , 得  $J = 0.11852 b^4$ .

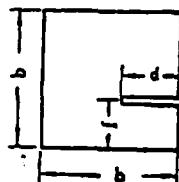


Table 2. Table of  $K_3$  Stress Intensity Factor  
 Coefficient  $B\left(\frac{d}{b}, \frac{f}{b}\right)$  of a Rectangular  
 Cross-section with Cracks.

/170

$d/b$  (0.05, 0.10, ..... 0.90)  
 $f/b$  (0.1, 0.2, 0.3, 0.4, 0.5)

$d/b$ $f/b$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.1	0.84847	1.0932	1.2558	1.3782	1.4733	1.5463	1.5998	1.6346	1.6501
0.2	1.3027	1.6779	1.9157	2.0941	2.2350	2.3487	2.4379	2.5022	2.5388
0.3	1.5869	2.0887	2.3803	2.6191	2.8155	2.9783	3.1182	3.2334	3.3121
0.4	1.7459	2.2909	2.6541	2.9415	3.1888	3.4044	3.5948	3.7573	3.8846
0.5	1.7845	2.3603	2.7480	3.0531	3.3178	3.6539	3.7658	3.9507	4.1094
$d/b$ $f/b$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.1	1.8454	1.6184	1.5569	1.4878	1.3778	1.2337	1.0524	.83282	.57933
0.2	2.5433	2.5102	2.4336	2.3070	2.1254	1.8855	1.5878	1.2353	.84208
0.3	3.3492	3.3378	3.2590	3.1028	2.8879	2.6463	2.1410	1.6696	1.1238
0.4	3.9646	3.9828	3.9261	3.7893	3.4994	3.1194	2.6279	2.0369	1.3781
0.5	4.2018	4.2381	4.1898	4.0377	3.7848	3.3828	2.8350	2.1976	1.4870

Equation:  $K_3 = B\left(\frac{d}{b}, \frac{f}{b}\right) \cdot M/b^{5/2}$ .

Example: 若  $\frac{d}{b} = 0.35$ ,  $\frac{f}{b} = 0.2$ , 得  $K_3 = 2.4379 M/b^{5/2}$ .

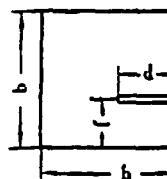


Table 3. Table of Torsional Rigidity Coefficient  
 $A\left(\frac{d}{b}, \frac{a}{b}\right)$  of Rectangular Cross-section  
 with Crack

/171

$d/b (0, 0.05, \dots, 0.95)$   
 $a/b (0.1, 0.2, 0.25, 0.5, 1, 2.5, 5)$

$\frac{d/b}{a/b}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	23.301	22.698	21.773	20.792	19.800	18.815	17.820	16.825	15.830	14.835
0.20	159.59	156.54	150.66	143.63	136.11	128.39	120.60	112.81	105.06	97.390
0.25	285.85	280.79	270.63	257.94	244.04	229.59	214.95	200.31	185.84	171.77
0.50	1405.8	1387.1	1344.9	1287.6	1221.1	1149.6	1078.4	1003.8	933.84	867.86
1.0	4573.7	4534.6	4446.9	4328.0	4179.8	4022.3	3857.6	3695.5	3545.0	3398.5
2.5	14587	14509	14400	14252	14077	13887	13691	13497	13322	13140
5.0	31245	31181	31037	30881	30702	30521	30326	30124	29943	29778
$\frac{d/b}{a/b}$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.10	13.841	12.847	11.855	10.867	9.8909	8.9377	8.0348	7.2359	6.6259	6.3032
0.20	89.849	82.499	75.428	68.747	62.620	57.229	52.797	49.527	47.535	46.718
0.25	158.04	144.91	132.58	121.27	111.37	102.83	96.208	91.504	88.892	87.880
0.50	807.11	752.55	704.92	664.74	632.26	607.50	590.07	579.22	573.72	571.89
1.0	3285.1	3150.8	3064.1	2975.2	2914.2	2869.7	2839.9	2882.5	2814.3	2813.7
2.5	12986	12862	12740	12649	12580	12630	12498	12480	12471	12469
5.0	29628	29498	29390	29304	29239	29193	29164	29147	29139	29137

Equation:  $D = \mu J, J = A\left(\frac{d}{b}, \frac{a}{b}\right) \cdot 10^{-4} \cdot b^4$ .

Example: 若  $\frac{d}{b} = 0.45, \frac{a}{b} = 1$ , 得  $J = 0.33985 b^4$ .

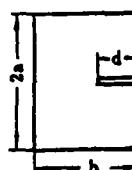


Table 4. Table of  $K_3$  Stress Intensity Factor Coefficient  
 $B\left(\frac{d}{b}, \frac{a}{b}\right)$  of Rectangular Cross-section with  
 Crack.  $d/b (0.05, 0.10, \dots, 0.90)$   
 $a/b (0.1, 0.2, 0.25, 0.5, 1, 2.5, 5)$

/172

$\frac{d/b}{a/b}$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	17.355	20.094	21.395	22.483	23.687	25.070	26.560	28.238	30.135
0.20	6.0942	7.5775	8.4240	9.0908	9.7208	10.371	11.074	11.847	12.694
0.25	4.4347	5.6166	6.3455	6.9182	7.4473	7.9792	8.5357	9.1081	9.7291
0.50	1.7945	2.3603	2.7460	3.0631	3.3178	3.6539	3.7858	2.9507	4.1004
1.0	.78944	1.0337	1.1989	1.3274	1.4134	1.4856	1.5250	1.5460	1.5424
2.5	.28278	.35231	.39894	.42930	.44759	.45659	.45628	.44863	.43409
5.0	.14628	.17062	.18746	.19801	.20369	.20503	.20273	.19722	.18887
$\frac{d/b}{a/b}$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.10	32.303	34.790	37.651	40.924	44.573	48.308	51.316	51.445	46.199
0.20	13.614	14.589	15.572	18.466	17.710	17.237	16.485	14.481	11.017
0.25	10.387	11.093	11.597	12.020	12.166	11.861	10.942	9.2492	8.7580
0.50	4.2018	4.2381	4.1898	4.0377	3.7848	3.3628	2.8350	2.1976	1.4870
1.0	1.5154	1.4548	1.3689	1.2539	1.1115	.94768	.76357	.56626	.36796
2.5	.40303	.38593	.35337	.31570	.27384	.22829	.18042	.13146	.083846
5.0	.17795	.16475	.14948	.13243	.11394	.094326	.074009	.053634	.034092

$$\text{Equation: } K_3 = B\left(\frac{d}{b}, \frac{a}{b}\right) \cdot M/b^{3/2}.$$

$$\text{Example: 若 } \frac{d}{b} = 0.75, \frac{a}{b} = 0.5, \text{ 得 } K_3 = 3.3628 M/b^{3/2}.$$

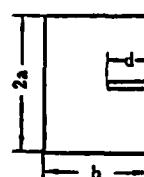


Table 5. Table of Torsional Rigidity Coefficient  
 $A\left(\frac{d}{b}, \frac{a}{2b}\right)$  of Rectangular Cross-section  
 with Crack

/173

$d/b (0, 0.05, \dots, 0.95)$   
 $a/2b (0.1, 0.25, 0.5, 1, 2.5, 5)$

$\frac{d/b}{a/2b}$	0	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	23.301	22.871	22.115	21.227	20.283	19.314	18.334	17.349	16.361	15.372
0.25	285.85	282.73	275.95	266.73	255.78	243.69	230.79	217.38	203.73	189.94
0.50	1405.8	1394.4	1388.8	1332.0	1288.1	1233.6	1176.1	1115.3	1052.8	990.10
1.0	4573.8	4548.6	4498.5	4420.8	4325.9	4215.8	4094.1	3984.8	3831.1	3696.2
2.5	14568	14523	14452	14367	14240	14107	13981	13806	13649	13486
5.0	31252	31179	31088	30974	30845	30704	30554	30397	30238	30081
$\frac{d/b}{a/2b}$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.10	14.382	13.393	12.405	11.420	10.441	9.4770	8.5441	7.6823	6.9430	6.4271
0.25	176.21	162.71	149.81	137.10	125.42	114.80	105.55	97.981	92.299	88.858
0.50	928.18	868.29	811.56	758.97	711.44	669.73	634.76	607.04	587.06	575.27
1.0	3563.1	3434.8	3313.4	3201.3	3100.5	3012.8	2939.6	2882.3	2841.7	2818.4
2.5	13329	13178	13036	12905	12789	12688	12608	12543	12600	12478
5.0	29928	29783	29648	29526	29419	29329	29256	29201	29164	29143

Equation:  $D = \mu J, J = A\left(\frac{d}{b}, \frac{a}{2b}\right) \cdot 10^{-4} (2b)^4$ .

Exmaple if:  $\frac{d}{b} = 0.35, \frac{a}{2b} = 2.5, J = 1.3806(2b)^4$ .

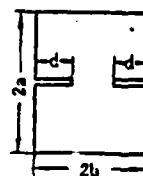


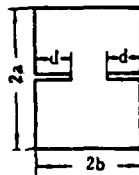
Table 6. Table of  $K_3$  Stress Intensity Factor Coefficient  
 $B\left(\frac{d}{b}, \frac{a}{2b}\right)$  of Rectangular Cross-section with  
Crack.

$d/b$  (0.05, 0.10,.....,0.90)  
 $a/2b$  (0.1, 0.25, 0.5, 1, 2.5, 5)

$\frac{d/b}{a/2b}$	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
0.10	15.176	18.404	20.206	21.599	22.894	24.221	25.650	27.228	28.997
0.25	3.5426	4.6061	5.3436	5.9542	6.5083	7.0463	7.5883	8.1527	8.7545
0.50	1.3858	1.8334	2.1690	2.4474	2.6970	2.9316	3.1547	3.3874	3.5718
1.0	.61301	.79714	.93640	1.0489	1.1438	1.2253	1.2952	1.3539	1.4010
2.5	.23460	.28260	.32088	.36134	.37510	.39305	.40599	.41417	.41776
5.0	.13067	.14580	.15832	.16831	.17598	.18144	.18474	.18594	.18612
$\frac{d/b}{a/2b}$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
0.10	31.002	33.291	35.919	38.939	42.369	46.114	49.649	51.951	50.420
0.25	9.3861	10.045	10.714	11.358	11.903	12.249	12.238	11.663	10.228
0.50	3.7841	3.9345	4.0729	4.1841	4.1900	4.1208	3.9267	3.6776	3.0183
1.0	1.4348	1.4545	1.4574	1.4396	1.3979	1.3286	1.2280	1.0830	.88685
2.5	.41676	.41110	.40047	.38488	.36387	.33650	.30212	.25952	.20711
5.0	.18237	.17767	.17088	.16200	.15093	.13784	.12202	.10382	.082569

Equation:  $K_3 = B\left(\frac{d}{b}, \frac{a}{2b}\right) \cdot M/(2b)^{5/2}$ .

Example if:  $\frac{d}{b} = 0.35, \frac{a}{2b} = 2.5, K_3 = 0.40599 M/(2b)^{5/2}$ .



#### REFERENCES

- [1] Westmann, R. A. and Yang, W. H. Stress Analysis of Cracked Rectangular Beams. *J. A. M.*, 34, 3, (1967), pp. 693-701.
- [2] Sih G. C. Strength of Stress Singularities at Crack Tips for Flexural and Torsional Problems. *J. A. M.*, 30, 3, (1963), pp. 419-425.
- [3] Timoshenko, S., The Theory of Elasticity, Chinese translation, (1964), pp. 272-277.
- [4] Qian Weizhang et al, The Theory of Elastic Cylinder torsion, (1958), pp. 22-40.
- [5] Wu Kinmou, Mathematical Physics Equations (Vol. 2), (1958), pp. 177.
- [6] Liebowitz, H., Fracture Vol. 2, (1968), pp. 231-232.
- [7] Арутюян, Н. Х.. Гулкянян, Н. О.. О центре изгиба некоторых призматических стержней с полигональным поперечным сечением. *ПММ*, 18, 5, (1954), 597-618.
- [8] Sih, G. C., Handbook of Stress-intensity Factors, (1973), pp. 2.5.1-3.
- [9] Chen Kuanzhou, Methods of calculating the torsions rigidity of rectangular section bar with crocus, unpublished.

## Summary

# The Automatic Matrix Force Method and Techniques for Handling More Complex Computations with Given Computer Capacity

Yang Qingxiong

Usually, the matrix displacement method, including the direct stiffness method, is used in aircraft structural analysis. But the automatic matrix force method, as presented in this paper, is sometimes to be preferred; computer programs for it were accordingly prepared and applied to the detail analysis of structural joints and to the integral analysis of major assemblies of an airplane.

In the early days of the matrix force method, both the selection of the basic system and the calculation of  $b_0$  and  $b_1$  matrices had to be done by hand. The manual method is very troublesome and mistakes are hard to avoid. For this reason, in the case of complex structures with high degree of redundancy, the matrix force method was abandoned. The automatic matrix force method presented in this paper can overcome the above-mentioned difficulties, and the basic system and the  $b_0$ ,  $b_1$  matrices can be obtained automatically in a straightforward manner without iteration. The procedure is as follows:

The equilibrium equations of all the structural elements are written out. The augmented matrix of these equations are eliminated by the Gauss-Jordan method. Then, the residual columns of the augmented matrix after elimination will automatically give the basic system and the  $b_0$ ,  $b_1$  matrices (a proof of this conclusion is given in this paper). With  $b_0$  and  $b_1$  matrices known, the regular procedure of the matrix force method is used, i. e., the flexibility matrix  $f$  and the applied load column vector  $P$  are substituted into the formula  $S = b_0 P - b_1 (b_1' f b_1)^{-1} b_1' f b_0 P$  to find the unknown internal force column vector  $S$ . In eliminating the augmented matrix, the largest element is always chosen as the pivot. In the special case where the chosen largest element happens to be zero, a discussion about the meaning and the proper treatment of such a special case is also given in this paper.

With a 32 K internal storage and a 24 K external storage, "program A" prepared for the procedure presented above can solve a problem of the following size:

number of equilibrium equations	$m \leq 155$ ,
number of unknown internal forces	$n \leq 180$ ,
number of applied loads	$n_p \leq 25$ ,
number of redundancy	$n_c \leq 87$ ,
with an additional condition	$n(n + n_p) \leq 27800$ .

In order to cope with problems of greater size, the sparse matrix method is used in this paper for storing and operating with the augmented matrix and the flexibility matrix. Furthermore, the symmetric and positive-definite property of the  $d_{11}(=b_1^T b_1)$  matrix is utilized to permit the "halfstore" operation in forming  $d_{11}$  matrix and in finding its inverse. With the same internal and external storage as before, "program B" can handle a problem of greater size as follows:

$m \leq 2000$ ,
$n \leq 2210$ ,
$n_p \leq 50$ ,
$n_c \leq 210$ ,

an additional condition being that the sum of nonzero elements of the  $b_1$  and  $b_1^T$  matrices is less than or equal to 12000.

THE AUTOMATIC MATRIX FORCE METHOD AND TECHNIQUES FOR HANDLING  
MORE COMPLEX COMPUTATIONS WITH GIVEN COMPUTER CAPACITY

Yang Qingxiong

Abstract

In this paper we reviewed the history of using matrix force method for structural computation in domestic aeronautic circles and the development and application of automatic matrix force methods in recent years.

In the early days the matrix force method was limited in its application because both the selection of the fundamental system and the computation of the matrices  $b_0$ ,  $b_1$  required hand-calculation. The automatic matrix force method was what brought a solution to this problem. We shall describe the principles of this method: For a statically indeterminate system when the equilibrium equations for all the elements are written down, the augmented matrix may then be eliminated by the Gauss-Jordon method and finally the fundamental system may be selected and the matrices  $b_0$ ,  $b_1$  obtained.

Problems solvable by the basic automatic matrix force method are very limited in size. By using the sparse matrix method we may increase the sizes of the augmented and the flexibility matrices. Also the symmetricity and positive definiteness of the  $d_{11}$  matrix may be used to take advantage of the "half store" operation in increasing the size of  $d_{11}$ . Combination of these two methods will greatly extend the size of the problem.

I. FORWARD

From very early on our domestic aeronautic design office

had made use of the first generation electronic computers to do structural analysis. The principal references at that time were the papers [1], [2] by Argyris. The matrix force method and the matrix displacement method were both used. Later on it was felt that for statically indeterminate systems of high rank with complex structure, hand calculation of  $b_0$  and  $b_1$  was difficult and error-prone and emphasis was gradually shifted toward the displacement method. (Argyris did the same thing and later used the displacement method exclusively). Thenceforth, the displacement method was practically the only method used in domestic aeronautic circles (including the stiffness method). However, because the force method had a long history of development and is more direct and comprehensible for structural strength computation and design, people continued to research it. As an example, Professor Chan Baipin spent years in research and finally developed a way to formulate the matrix force method from the equilibrium equations.

In recent years detailed analysis was required for fatigue lifetime computations and the force method was favored. According to foreign references the force method was also used [3]. The lack of reference resources on the method compelled us to work independently until finally we mastered the automatic matrix force method and applied it to practical problems such as the detailed analysis of rivet connections [4] and the structural analysis of the mid section of a wing. In fact, Denke et al. at Douglas in the U.S. have long mastered the automatic matrix force method (which they refer to as redundant force method). They started to code a general computer program for the automatic matrix force method in 1959, used it in 1967 and are now developing a large program called FORMAT encompassing both the force method and the displacement method [5]. At the same time, other people in U.S. and in Canada such as Robinson et al. [6] also worked on the automatic matrix force method. However, during this period, domestically we did not have any

material to introduce the works of Denke et al, and what we could find was all rather vague. As a consequence we failed to pay much attention to it. It was not until after 1974 that we obtained books [eg. 9,10] referring to this method without much detailed explanations.

In this paper we shall describe the principles of our own automatic matrix force method and the practical measures that we have used to extend the size of the computation.

## II. PRINCIPLES OF THE AUTOMATIC MATRIX FORCE METHOD

As we pointed out in the elementary matrix force method, the internal forces in a statically indeterminate system are the superposition of the internal forces of the fundamental system and the unknown force system

$$S = b_0 P + b_1 X \quad (1)$$

and the unknown force may be obtained from the normal equations reflecting the displacement compatibility conditions

$$b_1^T f b_1 X + b_1^T f b_0 P = 0 \quad (2)$$

where  $S$  = total internal force column vector

$P, X$  = external load and unknown force column vector

$b_0$  = transformation matrix from external load to fundamental system internal force

$b_1$  = transformation matrix from unknown force to internal force in the unknown force system

$f$  = flexibility matrix

$n$  - number of total internal forces

$n_p$  = number of external loads

$n_c$  = redundancy number i.e. number of unknown forces or number of cutouts.

$X$  may be solved from (2) and substituted into (1) to obtain

$$S = b_0 \ P - b_1 (b_1^T f \ b_1)^{-1} b_1^T f \ b_0 \ P \quad (3)$$

where  $P$  is known,  $f$  may be written down straight-forwardly and  $S$  may be computed if  $b_0$ ,  $b_1$  are found. The key to automatic matrix force method is to find  $b_0$  and  $b_1$  automatically. (In the FORMAT program this computational program is called "the structural disector".) This aim is to be realised by applying the Gauss-Jordon elimination method to the system of equilibrium equations

$$A \ S = B \ P \quad (4)$$

where  $A$  = coefficient matrix of internal forces

$B$  = transformation matrix from external load to the RHS term of the system of equilibrium equations.

The overall equation of equilibrium of the structure need not be included in the system of equilibrium equations since it is linearly dependent on the above equation.

To apply the Gauss-Jordon method to a square matrix  $A$  is to reduce it to a unit matrix  $E$ . Now the  $A$  in (4) is not a square matrix for  $m < n$ . Thus  $S$  cannot be solved directly. Nonetheless we still apply the Gauss-Jordon elimination method on it. To avoid getting zero for the principal element on the diagonal (thus bringing the computational process to a halt) as well as to reduce computational errors, we choose the major element (i.e. element with the largest absolute value) as the pivot element. The major element may be chosen by rows or it may be chosen from all the elements in the remaining rows and columns

not yet computed. We shall use the first method. The column being eliminated may not be in sequence when we eliminate the major element as the pivot element. In order to arrange the columns eliminated in sequence (which is really not necessary), we move the column with the major element forward so that the major element falls on the diagonal. After  $m$  eliminations, the first  $m$  columns of  $\mathbf{A}$  are reduced to  $\mathbf{E}$ , leaving  $n-m = n_c$  columns not yet reduced to unit column vectors. To maintain the validity of the RHS of equation (4), when the columns of the first matrix  $\mathbf{A}$  are interchanged, we also interchange the corresponding rows of  $\mathbf{S}$  so that their product remains unchanged. After  $m$  steps, (4) becomes

$$\mathbf{A}^{(m)} \mathbf{S} = \mathbf{B}^{(m)} \mathbf{P} \quad (5)$$

where the superscript  $m$  denotes that this is the result after  $m$  eliminations and the symbol  $\sim$  denotes that columns and their corresponding rows have been interchanged. In the computation it is a general practice to use the augmented matrix form, i.e. equation (4) and (5) are written as

$$[\mathbf{A} \mid \mathbf{B}]_{m \times n} \quad (4')$$

$$[\mathbf{A}^{(m)} \mid \mathbf{B}^{(m)}]_{m \times n} \quad (5')$$

Since the first  $m$  column of  $\mathbf{A}^{(m)}$  is an  $\mathbf{E}$ , we may separate it into 2 parts:

$$[\tilde{\mathbf{A}}_F^{(m)} \mid \tilde{\mathbf{A}}_R^{(m)} \mid \mathbf{B}^{(m)}]_{m \times n} \quad (5'')$$

also

$$\tilde{\mathbf{A}}_F^{(m)} = \mathbf{E}_{m \times m} \quad (6)$$

We may do the same thing to (4') and write

$$[\tilde{\mathbf{A}}_F \mid \tilde{\mathbf{A}}_R \mid \mathbf{B}]_{m \times n} \quad (4'')$$

Similarly we exchange corresponding rows in  $S$  and separate it so that (4) remains valid

$$[\mathbf{A}_F | \mathbf{A}_K] \begin{bmatrix} \mathbf{S}_F \\ \mathbf{S}_K \end{bmatrix} = \mathbf{B}, \quad P$$

After expanding, we get

$$\mathbf{A}_F \mathbf{S}_F + \mathbf{A}_K \mathbf{S}_K = \mathbf{B}, \quad P \quad (7)$$

We may represent the  $m$  step elimination process as a transformation matrix  $T$ , i.e.

$$[\mathbf{A}_F^{(m)} | \mathbf{A}_K^{(m)} | \mathbf{B}^{(m)}] = T [\mathbf{A}_F | \mathbf{A}_K | \mathbf{B}]$$

From this, then

$$\mathbf{A}_F^{(m)} = T \mathbf{A}_F$$

$$\mathbf{A}_K^{(m)} = T \mathbf{A}_K$$

$$\mathbf{B}^{(m)} = T \mathbf{B}$$

Comparing the first equation and equation (6), it may be seen that

$$T = \mathbf{A}_F^{-1} \quad (3.1)$$

Substituting this into the last two equations, then

$$\mathbf{A}_F^{(m)} = \mathbf{A}_F^{-1} \mathbf{A}_F \quad (8.2)$$

$$\mathbf{B}^{(m)} = \mathbf{A}_F^{-1} \mathbf{B}, \quad (8.3)$$

In the following we shall investigate the physical meaning of (5") in two cases:

(1) fundamental system and  $b_0$

We cut  $\tilde{S}_L$  at the  $n_c$  internal forces (the number of cutouts is exactly equal to the number of unknown forces  $n_c$ ), then

$$\tilde{S}_L = O \quad (9)$$

Equation (7) becomes

$$\tilde{A}_F \tilde{S}_U = B, P$$

$\tilde{A}_F$  is an  $m \times m$  square matrix. If  $\tilde{A}_F^{-1}$  exists, then

$$\tilde{S}_U = \tilde{A}_F^{-1} B, P$$

From (8.1) we know that if we can realise  $m$  steps in the elimination process (we will also discuss the case where this is not possible later), then  $\tilde{A}_F^{-1}$  actually exists and hence  $\tilde{S}_U$

has a set of unique solutions. For a statically indeterminate system under external load  $P$ , if we can obtain a set of unique solutions to the above equation after making  $n_c$  cutouts, then apparently the system with the cutouts is a fundamental system  $\tilde{S}_U$

Substituting (8.3) into the above equation yields

$$\tilde{S}_U = B^{(m)} P$$

Combining equation (9) into the above, then

$$\tilde{S}_L = \begin{bmatrix} \tilde{S}_U \\ \tilde{S}_L \end{bmatrix} = \begin{bmatrix} B^{(m)} P \\ O \end{bmatrix} = \begin{bmatrix} B^{(m)} \\ O \end{bmatrix} P$$

By definition,

$$\tilde{\mathbf{b}}_0 = \begin{bmatrix} \mathbf{B}^{(m)} \\ \mathbf{0} \end{bmatrix} \quad (10)$$

This is to say that if we augment  $\mathbf{B}^{(m)}$  in (5") with  $\mathbf{0}_{m \times n_p}$  after applying  $m$  step Guass-Jordon elimination method to (4"), then we shall obtain  $\tilde{\mathbf{b}}_0$  which needs to be row-exchanged into the right sequence.

To obtain  $\mathbf{b}_0$  with the original row sequence, all we need to do is reverse the row interchange process that obtains  $\mathbf{S}_1$  from  $\tilde{\mathbf{S}}_1$ .

## (2) Unknown force system and $\mathbf{b}_1$

When the external load is taken away, the unknown force system may be obtained by applying the unknown forces  $\mathbf{X}_{e \times 1}$  to the various cutouts of the fundamental system. Namely, when we substitute

$$\begin{aligned} \mathbf{P}_{e \times 1} &= \mathbf{0} \\ \tilde{\mathbf{S}}_L &= \mathbf{X}_{e \times 1} \end{aligned} \quad (11)$$

into equation (7), then

$$\mathbf{A}_F \tilde{\mathbf{S}}_U + \mathbf{A}_R \mathbf{X}_{e \times 1} = \mathbf{0}$$

Since  $\mathbf{A}_F^{-1}$  exists,  $\tilde{\mathbf{S}}_U$  may be solved and the relation between  $\tilde{\mathbf{S}}_U$  and  $\mathbf{X}$  may be derived, i.e.

$$\tilde{\mathbf{S}}_U = -\mathbf{A}_F^{-1} \mathbf{A}_R \mathbf{X}_{e \times 1}$$

We substitute (8.2) in this

$$\bar{S}_1 U = - \bar{A}_R^{(m)} X$$

and combine with (11)

$$\bar{S}_1 = \begin{bmatrix} \bar{S}_U \\ \bar{S}_L \end{bmatrix} = \begin{bmatrix} - \bar{A}_R^{(m)} X \\ \bar{X} \\ \bar{X}_1 \end{bmatrix} = \begin{bmatrix} - \bar{A}_R^{(m)} \\ - \bar{X} \\ \bar{E} \end{bmatrix} X_1$$

By definition, then

$$\bar{b}_1 = \begin{bmatrix} - \bar{A}_R^{(m)} \\ - \bar{X} \\ \bar{E} \end{bmatrix} \quad (12)$$

That is: if we add a negative sign to all the rows in (5") not yet eliminated and then augment it from below with  $\bar{E}$ , then we get  $b_1$  which needs to be row-exchanged into the right sequence.

During elimination, a special case may arise when the maximum element chosen is zero which terminates the process. The maximum element should have the largest absolute value. If it is zero, then all the elements in that row must vanish. This may be separated into two cases. (1) If the elements in all the augmented columns of this row are also zero, then this row is linearly dependent on the other rows above it and is not an independent equation. We may then eliminate it all together and continue with the computation. The final degree of redundancy is  $n_c = n - (m - \text{number of linearly dependent equations})$ . (2) If not all the elements of the augmented columns are zero, then this row is incompatible with the previous rows, and is a contradictory equation. Now if we skip over this row and continue with our computation, we can only check for more dependent or contradictory equations. The result is meaningless. The appearance of contradictory equations may be caused by errors in the current or previous rows of the equilibrium equations. If not, then it may be due to the fact that the structure in the stress

direction as represented by one of the equations is unable to support the load, i.e. it contains mechanical structures.

In general, by this elimination method, we can in one sweep (1) examine whether there exist dependent or contradictory equations and identify them (2) automatically find the fundamental system and (3) automatically find  $b_0$  and  $b_1$ .

In actual computation, it is not necessary to exchange rows after selecting the major element. We need only record the rule of the exchange and use it to exchange rows for  $b_0$  and  $b_1$ .

The computer program (program a) developed for this basic computation under the condition of 32K word internal memory and 24K word external memory can solve problems of the following size: number of equilibrium equations  $m \leq 155$ , number of internal forces  $n \leq 180$ , number of external loads  $n_p \leq 25$  (with the auxiliary condition  $m(n + n_p) \leq 27800$ ), degree of redundancy  $n_c \leq 80$  (may be extended to 87).

### III. METHOD TO EXTEND PROBLEM SIZE

The basic computational method above limits the problem size because (1) the augmented matrix of the equilibrium equations is too big. Its number of elements is  $m(n + n_p)$ . Maximum problem size is reached when it fills the internal memory. (2) The coefficient matrix  $d_{11} ( = b_1^T f b_1 )$  of the normal equations has a limited size since it has to be stored in the internal memory together with  $b_1$  and  $f$ . To solve these two problems, we have adopted the sparse matrix method and made use of the symmetric positive definiteness of  $d_{11}$ . Program b, developed in this way, with 32 K internal memory and 24K external memory, can solve problems with  $m \leq 2000$ ,  $n \leq 2210$ ,  $n_p \leq 50$  (with the auxiliary condition that the number of non-zero elements in  $b_0$  and  $b_1$  together may not exceed 12000),  $n_c \leq 210$ .

## A. Application of Sparse Matrix Method

We have applied the sparse matrix method in the following two situations:

### 1. Augmented Matrices

Although the size of the augmented matrix is very large, its non-zero elements in each row generally number between 2 to 5, with most of the elements zero. Therefore we have adopted the sparse matrix storage method -- storing only the non-zero elements and not the zero elements. The actual methods of storage, retrieval and counting may be found in another paper. We briefly introduce it as follows:

The non-zero elements of the augmented matrix are randomly distributed. For each non-zero element, both its row and column number must also be stored. Two memory locations are used to store data for a non-zero element, the first to store its value and the second to store its related information. We shall call the two-location unit the "address" of the element and assign to each address a sequence number. Thus the 2-dimensional array AXS(1:12000, 1:2) has its row number equal to the "address" number with a total of 12000 addresses. When storing, we take the non-zero elements of the augmented matrix by row and store it in AXS with the column number in increasing sequence. The first non-zero element of a row in the augmented matrix is called the "head" of that row and the last non-zero element the "tail" of that row. We then define another one dimensional integer array HTW (1:m) to store the "head" number HT and the "tail" number HW of each row, i.e. in HTW (i) we record the "head" number HT and the "tail" number HW of the  $i^{\text{th}}$  row in a compact way, using the rightmost 5 digits for HW and the 6th and higher digits for HT. Thus we can determine the position of each element if we store the column number in the second word of the "address". For

a particular row number  $i$  and column number  $j$  we first find from HTW ( $i$ ) the values for HT and HW, then we check the number LH in the second word of each of the "address" from HT to HW to see if it is equal to  $j$ . If so, then the value stored in the first word at that address must be the  $i$ ,  $j$ th element in the augmented matrix. Otherwise the value must be zero.

We assign 12000 addresses to AXS but when we store the augmented matrix in memory according to the sparse matrix method, usually only a small portion of the addresses is used. The last address number is called WH. From WH to the end address number 12000 is an empty "warehouse".

In the elimination process sometimes we may reduce an element to zero and sometimes we may change a zero element to non-zero. Generally speaking, the non-zero elements will be greatly increased. Thus we need to move many of the addresses back and forth between the "warehouse" and the address array. To avoid this back and forth movement, we define an "artificial sequential chain". This is done by storing another integer -- the "next address" XYDZ in the second word of the "address", which shows the address number of the next address in the artificial sequential chain. To change the position of an address, we only need to change its artificial sequential chain number.

## 2. Flexibility matrix

The flexibility matrix is also a sparse matrix, but it is also a pseudodiagonal matrix with minor matrices arranged along the diagonal in a regular manner. Corresponding to the uniformly varying axial force moment is a  $2 \times 2$  minor matrix

$$\begin{pmatrix} \frac{1}{3} \left( \frac{I}{EA} \right), \frac{1}{6} \left( \frac{I}{EA} \right), \\ \frac{1}{6} \left( \frac{I}{EA} \right), \frac{1}{3} \left( \frac{I}{EA} \right), \end{pmatrix} = \frac{1}{6} \left( \frac{I}{EA} \right), \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \quad (13)$$

Corresponding to the support and reaction force moments is the  $1 \times 1$  minor matrix  $(\frac{1}{E A})$ , and corresponding to the shear plate (rectangle or nearly rectangular trapezoid) is a  $1 \times 1$  minor matrix  $(\frac{F}{G t})$ , etc. Using the typical sparse storage method for a band matrix, we "squeeze" the pseudo-diagonal matrix into a "column". Although this is really a sequence of minor matrix and not a true column, we arrange the similar elements together, factor out the common factor of the minor and only store the special individual factors. (e.g. in equation 13 we do not store  $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$  but only store  $\frac{1}{6} \left( \frac{1}{E A} \right)$ ). Thus we actually "squeeze" the matrix into a column. For example, the original  $f$  is  $m \times n$

$$\begin{bmatrix} \frac{1}{3} \left( \frac{l}{E A} \right), \frac{1}{6} \left( \frac{l}{E A} \right), \\ \frac{1}{6} \left( \frac{l}{E A} \right), \frac{1}{3} \left( \frac{l}{E A} \right), \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{3} \left( \frac{l}{E A} \right), \frac{1}{6} \left( \frac{l}{E A} \right), \\ \frac{1}{6} \left( \frac{l}{E A} \right), \frac{1}{3} \left( \frac{l}{E A} \right), \end{bmatrix}$$

$f =$

$$\left( \frac{l}{E A} \right)_{\dots}$$

$$\left( \frac{F}{G t} \right)_{\dots \dots}$$

$$= \frac{1}{6 E A} \left[ \begin{array}{c} \left[ \begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix} \right] f_1 \\ \left[ \begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix} \right] f_2 \\ \ddots \\ 6f_{\dots} \\ \ddots \\ 6f_{\dots \dots} \\ \ddots \end{array} \right]$$

(14)

After "squeezing", we store it as

$$\frac{1}{6E_0} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{u+1} \\ \vdots \\ f_{u+v+1} \end{bmatrix} \quad (15)$$

where

$$f_i = \left( \frac{l}{EA} \right)_i, \quad i = 1, 2, \dots, u$$

$$f_i = \left( \frac{l}{EA} \right)_i, \quad i = u+1, \dots, u+v$$

$$f_i = \left( \frac{F}{Gt} \right)_i, \quad i = u+v+1, \dots, u+v+w$$

$u, v, w$  = number of uniformly varying axial force moments, support and reaction force moments and shear plate respectively

$E_0$  = elastic modulus of the material chosen as standard modulus

$E_s = \frac{E_0}{E_s}$  = specific elastic modulus

$G_s = \frac{G_0}{E_0}$  = specific shear modulus

$E_i, G_i$  = elastic modulus and shear modulus of each element

$l, A, t, F$  = beam length, beam cross-section area, plate thickness, plate surface area

During computation, for any  $f_i$  the minor matrix must be at the  $i$ th row and  $i$ th column. If  $i \leq u$ , then we have to multiply  $f_i$  with  $\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$  while for  $u < i \leq u+v$  and  $u+v < i \leq u+v+w$ , we have to multiply  $f_i$  by 6. In this way an  $n \times n$  matrix need only be stored as a one dimensional array of size  $(u+v+w) \times 1$  with  $u+v+w \leq n$ , which is generally less than one row of the original  $f$  matrix.

B. Using the Symmetricity and Positive Definiteness of  $d_{11}$

The  $f$  matrix in the coefficient matrix of the force method normal equations

$$\frac{d_{11}}{n_c \times n_c} = \frac{b_1'}{n_c \times n_c} f \frac{b_1}{n_c \times n_c} \quad (16)$$

as in (14) is a pseudo-diagonal matrix. Since each of its minor matrices is positive definite and symmetric, the  $f$  matrix must itself be positive definite and symmetric. If the rank of  $\frac{b_1}{n_c \times n_c}$  is equal to  $n_c$ , then when  $f$  is symmetric positive definite,  $\frac{d_{11}}{n_c \times n_c}$  will also be symmetric positive definite. This may be proved briefly as follows. Multiply equation (16) on both sides with an arbitrary non-zero column vector  $\frac{a}{n_c \times 1}$  and its transpose

$$\begin{aligned} \frac{a'}{n_c \times 1} \frac{d_{11}}{n_c \times n_c} \frac{a}{n_c \times 1} &= \frac{a'}{n_c \times 1} \frac{b_1'}{n_c \times n_c} f \frac{b_1}{n_c \times n_c} \frac{a}{n_c \times 1} \\ &= (\frac{b_1}{n_c \times n_c} \frac{a}{n_c \times 1})' f (\frac{b_1}{n_c \times n_c} \frac{a}{n_c \times 1}) > 0 \end{aligned}$$

Since the rank of  $\frac{b_1}{n_c \times n_c}$  is  $n_c$ ,  $b_1$  must have  $n_c$  linearly independent columns. Since  $\frac{a}{n_c \times 1} \neq 0$ , then  $\frac{\beta}{n_c \times 1} = \frac{b_1}{n_c \times n_c} \frac{a}{n_c \times 1} \neq 0$  and  $\beta$  must also be an arbitrary non-zero column vector. According to an equivalent definition of symmetric positive definiteness,  $f$  symmetric is positive definite implies that  $\beta' f \beta > 0$ . From the above equation  $a' d_{11} a (= \beta' f \beta) > 0$ . Hence  $d_{11}$  is symmetric positive definite by reverse argument.

If some of the elements in the structure are stiff (e.g. the support and reaction force moment which represents the stiff support), then the corresponding  $f_{ij}$  in the flexibility matrix  $f$  will be zero. Thus  $f$  will be semi-positive definite and not positive definite. But the rows and columns corresponding to the vanishing  $f_{ij}$ 's are all zero, and, therefore, may be eliminated to reduce  $f$  to an  $(n-z) \times (n-z)$  matrix (assuming the

the number of vanishing rows to be  $z$ ). Likewise the corresponding rows of  $b_1$  may be eliminated to reduce its size to  $(n-z) \times n_c$ . From the above, we can see that provided  $n_c \leq n-z$ , the columns of the reduced  $b_1^*$  will remain linearly independent (which is always true for a correctly obtained fundamental system). Thus  $b_1^*$ 's rank is still  $n_c$  and  $d_{11}^*$  will still be symmetric positive definite. Hence, in general, even if there are still elements in the structure, provided that their number is not too large,  $d_{11}$  will always be symmetric and positive definite. A symmetric, positive definite matrix such as  $d_{11}$  may always be separated into a triangular matrix and its transpose, i.e.

$$\begin{matrix} d_{11} \\ n_c \times n_c \end{matrix} = \begin{matrix} L \\ n_c \times n_c \end{matrix} \begin{matrix} U \\ n_c \times n_c \end{matrix} \quad \text{and} \quad U = L'$$

where  $L$  is a lower triangular matrix and  $U$  an upper triangular matrix.  $U$  can be easily obtained if  $L$  is known. Thus we need only store  $L$  and not  $U$ . Since  $L$  is a lower triangular matrix, we need not store the zero elements. Only the non-zero elements in the lower triangle (including the diagonal elements) need be stored. As  $d_{11}$  is symmetric, we also need only store its lower triangular elements. Thus  $d_{11}$  may be "half-stored" and separated "at the same location" into the lower triangular part of  $L$ . In this way we only need  $n_c(n_c+1)/2$  memory locations, realizing a saving of nearly a half.

If  $d_{11}$  is obtained from (16), then we need at least to store in the internal memory the matrix  $b_1$ , the "squeezed"  $f$  and leave space for half-storing  $d_{11}$ . To save internal storage, we shall proceed as follows:

$f$  (or the reduced  $f^*$ ) is also a symmetric, positive definite matrix, and may be analysed as shown above, namely:

$$\begin{matrix} f \\ n_c \times n_c \end{matrix} = \begin{matrix} f_0 \\ n_c \times n_c \end{matrix} \begin{matrix} f_1 \\ n_c \times n_c \end{matrix},$$

where  $f_x$  is a lower triangular matrix. Since  $f$  is a pseudo-diagonal matrix, then  $f_x$  must also be a psuedo-diagonal matrix formed from lower triangular minor matrices derived from each minor matrix. For  $f$  in (14), the  $f_x$  is

$$f_x = \frac{1}{\sqrt{6}E_0} \begin{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} \end{bmatrix} \sqrt{f_1} \\ \begin{bmatrix} \sqrt{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} \end{bmatrix} \sqrt{f_2} \\ \vdots \\ \begin{bmatrix} \sqrt{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} \end{bmatrix} \sqrt{f_{u+1}} \\ \vdots \\ \begin{bmatrix} \sqrt{2} & 0 \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{6}}{2} \end{bmatrix} \sqrt{f_{u+v+1}} \\ \vdots \end{bmatrix}$$

The actual formation of  $f_x$  is not necessary. In our computation, we need only remember the form of  $f_x$  and store it in the form of (15) according to the method described in the passages after equation (15).

Once  $f_x$  is obtained, we may write  $d_{11}$  as

$$d_{11} = b_1' f_x f_1' b_1 = \underbrace{\begin{bmatrix} b_1' & f_x & f_1' \\ n_c \times n & n \times n & n \times n \end{bmatrix}}_{n_c \times n} \underbrace{\begin{bmatrix} b_1' f_x \\ n \times n \end{bmatrix}}_{n \times n}$$

As  $b_1'$  is an  $n_c \times n$  matrix, so is  $b_1' \times f_x$ . We shall form  $b_1' f_x$  from  $b_1'$  "at the same location" without using extra internal storage. We then compute the lower triangular part of  $d_{11}$  row by row from bottom up. After each row is computed, the corresponding row in  $b_1' f_x$  will no longer be needed. We may erase it and release the memory location for other use. Thus within the space for the "half-stored"  $d_{11}$ , we may store  $b_1'$  and the "squeezed"  $f$ , complete the multiplication of  $b_1' f_x$ , form the lower triangular part of  $d_{11}$  and separate  $d_{11}$  to obtain the lower triangular part of  $L$ .

The internal forces of the unknown force system are abbreviated as

$$S_1 = b_1 \begin{smallmatrix} X \\ \text{nx1} \end{smallmatrix} \quad X = -b_1 \begin{smallmatrix} b_1^T & f & b_1 \\ \text{nxn} & \text{nxn} & \text{nxn} \end{smallmatrix} \begin{smallmatrix} b_1^T & f & b_0 \\ \text{nxn} & \text{nxn} & \text{nxn} \end{smallmatrix} \begin{smallmatrix} P \\ \text{nx1} \end{smallmatrix}$$

$$S_1 = -b_1 \begin{smallmatrix} d_{11}^{-1} & d_{10} \\ \text{nxn} & \text{nxn} \end{smallmatrix} \begin{smallmatrix} P \\ \text{nx1} \end{smallmatrix} = -b_1 \begin{smallmatrix} d_{11}^{-1} & D_{10} \\ \text{nxn} & \text{nxn} \end{smallmatrix}$$

We shall not find the inverse of  $d_{11}$ , but shall use the fact that  $d_{11}$  has been separated to solve the equations for  $X$ , namely,

$$\begin{aligned} X &= -d_{11}^{-1} D_{10} = -\begin{smallmatrix} L & L' \\ \text{nxn} & \text{nxn} \end{smallmatrix}^{-1} \begin{smallmatrix} D_{10} \\ \text{nx1} \end{smallmatrix} \\ &= -\begin{smallmatrix} (L')^{-1} & L^{-1} \\ \text{nxn} & \text{nxn} \end{smallmatrix} \begin{smallmatrix} D_{10} \\ \text{nx1} \end{smallmatrix} \end{aligned}$$

Let  $\begin{smallmatrix} Y = L^{-1} & D_{10} \\ \text{nx1} & \text{nxn} \end{smallmatrix}$ , then  $\begin{smallmatrix} L & Y = D_{10} \\ \text{nxn} & \text{nx1} \end{smallmatrix}$ .  $Y$  may be conveniently solved from this set of equations since  $L$  is a lower triangular matrix. After  $Y$  is solved, we substitute it into the above equation to get  $\begin{smallmatrix} X = -(L')^{-1} Y \\ \text{nx1} \end{smallmatrix}$  or  $\begin{smallmatrix} L' & X = -Y \\ \text{nxn} & \text{nx1} \end{smallmatrix}$ .  $X$  is then obtained by solving the upper matrix equation and finally substitution of  $X$  into (1) will yield the final result -- the internal forces of all the elements  $S$ .

$\text{nx1}$

In conclusion, after applying the "half-store" methods as described above, with 32K internal memory and 24K external memory, we increased the degree of redundancy to 210 while in the original "full store" method (Prog. a) it was only 80 (at most 87).

## REFERENCES

- [1] Argyris, J.H., Energy Theorems and Structural Analysis. Part I. General Theory, Aircraft Engineering, Vol. 26 (1954), No. 10, p. 347, No. 11, p. 383; Vol. 27 (1955), No. 2, p. 42, No. 3, p. 80, No. 4, p. 125, No. 5, p. 145.
- [2] Argyris, J.H., and Kelsey, S., The Matrix Force Method of Structural Analysis and Some New Applications. ARC R&M 3034 (1956).
- [3] Deness, G.V., Fatigue Prediction Study, AD 273894, 即 WADD TR61-153 (1962).
- [4] Si Erjian: Estimating Fatigue Life of Aircraft Structure Connecting Parts, Unpublished.
- [5] Warren, D.S., Application Experience with the FORMAT Computer Program, Proceedings of the second Conference on Matrix Methods in Structural Mechanics (Held at Wright-Patterson Air Force Base, Ohio, 1968), AD 703685 or AFFDL-TR-68-150 (1969), pp. 839~867.
- [6] Robinson, J., Integrated Theory of Finite Element Methods, John Wiley & Sons (1973).
- [7] Denke, P.H., A General Digital Computer Analysis of Statically Indeterminate Structures, NASA TN D-1666 (1962).
- [8] Denke, P.H., Matrix Methods of Aerospace Strutural Analysis, Proceedings of the Second Conference on Matrix Methods in Structural Mechanics (Held at Wright-Patterson Air Force Base, Ohio, 1968), AD 703685 or AFFDL-TR-68-150 (1969), pp. 15~79.
- [9] J. S. Puchiminschy [as phoneticized], translated by Wang Derong et al. The theory of structural matrix analysis. Defense Industry Publishing House (1974).
- [10] Northern University of Communications, Railroad Construction Department. Structural matrix analysis. Chinese Construction Industry Publishing House (1974).

## Summary

# Structural Analysis of Fuselages with Cutouts by Finite Element Method

*Ge Shoulian, Sun Can, Tang Xuanchun,  
and Ye Tianqi*

Much investigation has been done in the analysis of fuselage structures with cutout, such as by Kuhn, Schletchte, ВЛЯСОВ, КЛИМОВ, and РУДЫХ. But the theories and techniques proposed in these papers are not suitable to the analysis of conical fuselage structures with large cutout; moreover the elastic behavior of bulkheads needs to be considered. A finite element method, which is different from Argyris's force method, is used in this paper to analyze fuselage structure with cutout. The method and the program developed can be applied to any of these fuselage structures.

As an example, the rear part of a transport aeroplane's fuselage is calculated. It consists of two sections, the front section being a closed cylinder with a floor and the other section containing a large cutout. An idealized structure is considered to have 1950 degrees of freedom. In this analysis, three kinds of finite element have been adopted. The first is a flange element with various cross sections. The second is a constant-shear-flow quadrilateral of arbitrary shape. The last is a beam element of various cross-section taking account of eccentricity of the nodes from the axes of moment of inertia. A factorization algorithm is adopted in the solution of the system of linear algebraic equations. The bandwidth of the sparse matrix is 396. A block decomposition and elimination procedure is included in the program. Packed form of storage in the computers has also been utilized. Calculations were completed in a computer with small storage.

According to the Wagner-ВЛЯСОВ theory, the tensile and compressive normal stresses in the fuselage have a two-wave circumferential variation. However, when the elastic behavior of the bulkhead is considered, the normal stresses have a three-wave circumferential variation (figure 3-5). This has been verified experimentally. The calculated displacements along the length of fuselage are in agreement with experimental data obtained in the full scale test of the aeroplane.

STRUCTURAL ANALYSIS OF FUSELAGE WITH CUTOUTS BY FINITE ELEMENT METHOD

Ge Shoulian, Sun Can, Tang Xuanchun, Ye Tianqi

In the past there have been many theoretical computational methods on the structure of fuselage with cutouts, but they all have their limitations. The rear section of the fuselage with the cutout as shown in Figure 1.1 possesses, on the one hand, large taper and conicity and, on the other hand, a very complicated internal structure. In analysing the structure, we have to compute simultaneously the middle section with a floor and the tail section with full body plate frame, both of which are closed sections. In addition, the stiff frame assumption is no longer appropriate, and we must take into consideration the effect of elasticity. Thus various traditional methods of theoretical computation cannot be used to solve this kind of structural problem. In this paper the finite element method is used in our analysis.

The method and program developed in our paper is suitable not only for a fuselage with cutout, but also for strength computation of a fuselage, wing or other parts. It can also be applied without any substantial difficulty to treat composite structures of diverse elements simply by suitably increasing the types of elements used.

In this paper we have chosen to use a constant shear stress plate element, eccentric beam element with variable cross-section and equal axial force bar element with variable cross-section. The capacity of the plate to bear normal stress is converted into the long trusses and the flanges. Because of the chain structure, the stiffness equations have a band form. The solution is obtained directly through factorization. The

singularity of the matrix is treated with the row and column reduction method. The dimension of the matrix is 1950 with a half bandwidth of 280. After compression, the dimension is 1239 with a half bandwidth of 198. The complete computation was carried out on a DJS-8 computer, and the results were in good agreement with the test data.

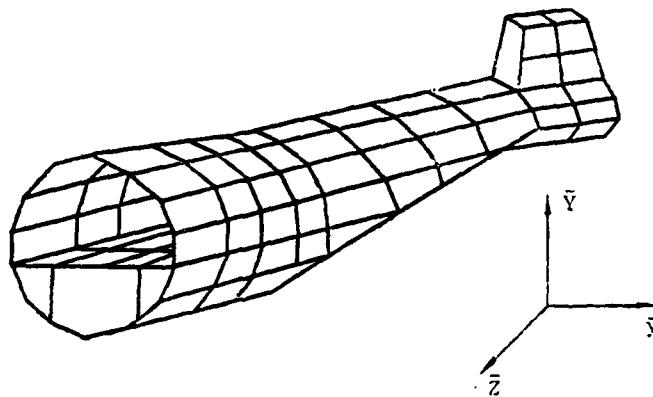


Figure 1.1. Tubular Model of the Rear Fuselage

### I. Idealization of the Structure

To simplify the practical structure into a computational model, the long trusses and the frames are combined. Generally, strengthened frames and long trusses are used in their computation, and the number of computed frames and trusses around the cutout region are suitably increased. The rigid support of the complete rear fuselage is in the mid section and the axial displacement of the nodal point in the plane of support where the long beam is cut is not zero. In the idealised model, the force on the floor of the luggage compartment is taken into account, while that on the floor of the fuel tank is not.

The dividing frame and the longitudinal and cross beams

of the floor in the structure are simplified as eccentric beam elements, or a combination of plate and bar elements. The former is used to simplify beams of small height, while the latter is used to simplify composite thin-walled beams of greater heights.

### III. Element Stiffness Matrix $[K]^e$

#### 1. Stiffness matrix of shearing plate element

The stiffness matrix of an arbitrary quadrilateral shearing plate element is derived with the plane coordinate system in Figure (2.1) as the local coordinate system. The shear stress  $\tau$  of the shearing plate element is constant. Hence we may consider the arbitrary quadrilateral shear plate as part of the rectangular pure shear plate as shown in Figure (2.2). The interaction between the quadrilateral shearing plate element 1234 and the rest of the rectangular pure shear plate is expressed through statically equivalent nodal point forces. For example, the nodal point force of nodal point 2 in the  $x$  direction is where  $t$  is the plate thickness. In Figure (2.2) we

$$F_{2x} = -\frac{1}{2} \tau t (x_2 - x_1) - \frac{1}{2} \tau t (x_3 - x_2) \quad (2.1)$$

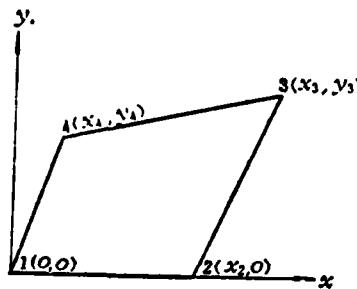


Figure 2.1

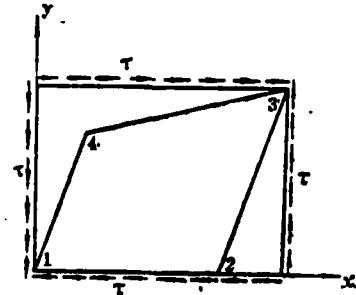


Figure 2.2

put nodal point 1 at the origin of the local coordinate system, so that  $x_1=0$ ,  $y_1=0$  and equation (2.1) is simplified to

$$F_{2x} = -\frac{1}{2} \tau t x_3 \quad (2.2)$$

The nodal point force at node 2 in the y direction is

$$F_{2y} = \frac{1}{2} \tau t (y_3 - y_2) = \frac{1}{2} \tau t v_3 \quad (2.3)$$

The equivalent nodal point forces of other nodes can be similarly found to obtain finally

$$\{F\} = \begin{pmatrix} F_{1x} \\ F_{1y} \\ F_{2x} \\ F_{2y} \\ F_{3x} \\ F_{3y} \\ F_{4x} \\ F_{4y} \end{pmatrix} = \frac{1}{2} \tau t \begin{pmatrix} x_4 - x_2 \\ -y_4 \\ -x_3 \\ y_3 \\ -(x_4 - x_2) \\ y_4 \\ x_3 \\ -y_3 \end{pmatrix} = \tau \{A\} \quad (2.4)$$

where  $\{A\} = \frac{1}{2} t \{ x_4 - x_2, -y_4, -x_3, -(x_4 - x_2), y_4, x_3, -y_3 \}$ . The nodal displacement corresponding to the nodal force is  $\{\delta\}$  where

$$\{\delta\} = \{u_1, v_1, u_2, v_2, u_3, v_3, u_4, v_4\} \quad (2.5)$$

in which  $u$  is the displacement along  $x$ ;  $v$  - the displacement along  $y$ ; and the subscript denotes the nodal point number.

According to the relation that the external work is equal to the

strain energy of the plate element, we have

$$\frac{1}{2} \{F\}^T \{\delta\} = \frac{1}{2} \frac{St}{G} \tau^2 \quad (2.6)$$

where  $S$  is the area of the arbitrary quadrilateral plate element and  $G$  is the shear modulus of the plate material.

Substituting equation (2.4) into (2.6), then

$$\tau = \frac{G}{St} \{A\}^T \{\delta\} \quad (2.7)$$

$$\tau \{A\} = \frac{G}{St} \{A\} \{A\}^T \{\delta\}$$

$$\{F\} = \frac{G}{St} \{A\} \{A\}^T \{\delta\}$$

Therefore the stiffness matrix of the arbitrary quadrilateral plate element is

$$[K]^e = \frac{G}{St} \{A\} \{A\}^T \quad (2.8)$$

i.e.

$$[K]^e = \frac{G}{4S} \begin{bmatrix} (x_4 - x_2) \\ -y_4 \\ -x_3 \\ y_3 \\ -(x_4 - x_2) \end{bmatrix} \begin{bmatrix} (x_4 - x_2), -y_4, -x_3, y_3, -(x_4 - x_2), y_4, x_3, -y_3 \end{bmatrix} \quad (2.9)$$

$$\begin{bmatrix} y_4 \\ x_3 \\ -y_3 \end{bmatrix} \quad (2.9)$$

where

$$2S = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} + \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_3 & y_3 \\ 1 & x_4 & y_4 \end{vmatrix}$$

$$\therefore x_1 = y_1 = y_2 = 0$$

$$\therefore 2S = x_3 y_4 + y_3 (x_2 - x_4)$$

(2.10)

Coordinate transformation:

The coordinates of the nodal point in the global coordinate system are  $\bar{X}$ ,  $\bar{Y}$ ,  $\bar{Z}$ . The direction cosines of the local  $x$ ,  $y$  axes relative to the global coordinate axes may be expressed by using the coordinates of the points 1, 2 and 3. If the direction cosines of the  $x$  axis and the  $y$  axis relative to the global axis are respectively  $l_1$ ,  $m_1$ ,  $n_1$  and  $l_2$ ,  $m_2$ ,  $n_2$ , then

$$\left. \begin{aligned} l_1 &= \frac{\bar{X}_2 - \bar{X}_1}{L_{12}} \\ m_1 &= \frac{\bar{Y}_2 - \bar{Y}_1}{L_{12}} \\ n_1 &= \frac{\bar{Z}_2 - \bar{Z}_1}{L_{12}} \end{aligned} \right\} \quad (2.11)$$

where

$$L_{12} = \left[ (\bar{X}_2 - \bar{X}_1)^2 + (\bar{Y}_2 - \bar{Y}_1)^2 + (\bar{Z}_2 - \bar{Z}_1)^2 \right]^{\frac{1}{2}} \quad (2.12)$$

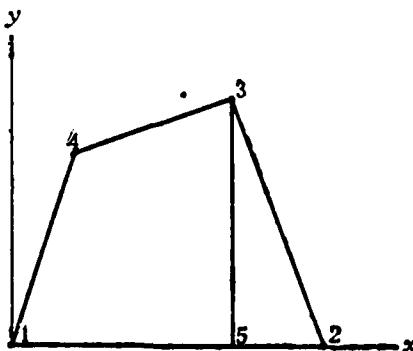


Figure 2.3.

Through point 3 draw a line perpendicular to the x axis intersecting it at point 5, as shown in Figure (2.3). Then

$$L_{15} = l_1(\bar{X}_3 - \bar{X}_1) + m_1(\bar{Y}_3 - \bar{Y}_1) + n_1(\bar{Z}_3 - \bar{Z}_1)$$

$$L_{ss} = \left[ (\bar{X}_3 - \bar{X}_1)^2 + (\bar{Y}_3 - \bar{Y}_1)^2 + (\bar{Z}_3 - \bar{Z}_1)^2 - L_{15}^2 \right]^{\frac{1}{2}} \quad (2.13)$$

The direction cosines are then

$$l_2 = \frac{[(\bar{X}_3 - \bar{X}_1) - l_1 L_{15}]}{L_{ss}}$$

$$m_2 = \frac{[(\bar{Y}_3 - \bar{Y}_1) - m_1 L_{15}]}{L_{ss}}$$

$$n_2 = \frac{[(\bar{Z}_3 - \bar{Z}_1) - n_1 L_{15}]}{L_{ss}} \quad (2.14)$$

Transformed to the global coordinate system, the stiffness matrix of the shearing plate element is

$$[\bar{K}]^e = [T]^T [K]^e [T] \quad (2.15)$$

$$[T] = \begin{bmatrix} [L] & & & \\ & [L] & & \\ & & [L] & \\ & & & [L] \end{bmatrix} \quad (2.16)$$

$$[L] = \begin{bmatrix} l_1 & m_1 & n_1 \\ l_2 & m_2 & n_2 \end{bmatrix} \quad (2.17)$$

From equation (2.9) we can see that  $[K]^e$  is expressed in terms of the nodal point coordinate values in the local coordinate system, but in structural calculations it is usually the global coordinate values of the nodal point that are known. For convenience, it is necessary to transform the local coordinate values in the stiffness matrix to global coordinate values. Their relations are

$$\begin{Bmatrix} x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{bmatrix} \bar{X}_2 - \bar{X}_1 & \bar{Y}_2 - \bar{Y}_1 & \bar{Z}_2 - \bar{Z}_1 \\ \bar{X}_3 - \bar{X}_1 & \bar{Y}_3 - \bar{Y}_1 & \bar{Z}_3 - \bar{Z}_1 \\ \bar{X}_4 - \bar{X}_1 & \bar{Y}_4 - \bar{Y}_1 & \bar{Z}_4 - \bar{Z}_1 \end{bmatrix} \begin{Bmatrix} l_1 \\ m_1 \\ n_1 \end{Bmatrix} \quad (2.18)$$

$$\begin{Bmatrix} y_3 \\ y_4 \end{Bmatrix} = \begin{bmatrix} \bar{x}_3 - \bar{x}_1, \bar{y}_3 - \bar{y}_1, \bar{z}_3 - \bar{z}_1 \\ \bar{x}_4 - \bar{x}_1, \bar{y}_4 - \bar{y}_1, \bar{z}_4 - \bar{z}_1 \end{bmatrix} \begin{Bmatrix} l_2 \\ m_2 \\ n_2 \end{Bmatrix} \quad (2.19)$$

The displacement column matrix in the global system is

$$\{\bar{\delta}\} = \{\bar{u}_1, \bar{v}_1, \bar{w}_1, \bar{u}_2, \bar{v}_2, \bar{w}_2, \bar{u}_3, \bar{v}_3, \bar{w}_3, \bar{u}_4, \bar{v}_4, \bar{w}_4\} \quad (2.20)$$

According to equation (2.7), the shear stress of the shearing plate element is, after being transformed from the local coordinate system to the global system,

$$\tau = \frac{G}{S_f} \{A\}^T [T] \{\bar{\delta}\} \quad (2.21)$$

where the local coordinate values in  $\{A\}^T$  are again transformed into their global values in accordance with (2.18) and (2.19).

## 2. Stiffness matrix of the beam element with variable cross-section

Figure (2.4) shows the beam element with variable cross-section in which  $EI$  changes linearly, i.e.

$$EI = EI_1 - \frac{x}{L} (EI_1 - EI_2) \quad (2.22)$$

Using the same method as applied to a beam with constant cross-section, we can derive the stiffness matrix of the beam element with variable cross-section as

$$\begin{Bmatrix} Q_2 \\ M_2 \\ Q_1 \\ M_1 \end{Bmatrix} = J \begin{bmatrix} C_{11} & & & \\ C_{12} - LC_{11} & L^2 C_{11} - 2LC_{12} + C_{22} & & \\ -C_{11} & LC_{11} - C_{12} & C_{11} & \\ -C_{12} & LC_{21} - C_{22} & C_{12} & C_{22} \end{bmatrix} \begin{Bmatrix} v_2 \\ \theta_2 \\ v_1 \\ \theta_1 \end{Bmatrix} \quad (2.23)$$

where

$$J = \frac{EI_2}{LD} \quad C_{11} = \frac{\ln P}{P-1} \quad C_{12} = \frac{L}{(P-1)^2} (1 - P + P \ln P)$$

$$C_{21} = C_{12} \quad C_{22} = \frac{L^2}{(P-1)^3} \left( P^2 \ln P - \frac{3}{2} P^2 + 2P - \frac{1}{2} \right) \quad P = \frac{EI_1}{EI_2}$$

$$D = \begin{vmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{vmatrix} = C_{11}C_{22} - C_{12}C_{21}$$

The bar element stiffness matrix with variable cross-section in which the bar cross-section changes linearly as shown in Figure (2.4) is

$$A = A_1 - \frac{x}{L} (A_1 - A_2) \quad (2.24)$$

Using a similar method as applied to a bar element with constant cross-section, we may derive the stiffness equations for a bar element with variable cross-section as

$$\begin{Bmatrix} F_1 \\ F_2 \end{Bmatrix} = \frac{EA_1}{L} \frac{(r-1)}{\ln r} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \end{Bmatrix} \quad (2.25)$$

where

$$r = \frac{EA_2}{EA_1}$$

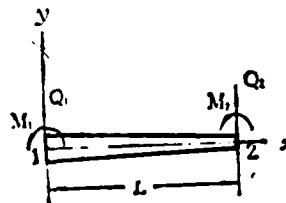


Figure 2.4

The stiffness equations of a beam element with variable cross-section under an axial load may be obtained by combining equation (2.23) and (2.25):

$$\begin{bmatrix} F_1 \\ Q_2 \\ M_2 \\ F_1 \\ Q_1 \\ M_1 \end{bmatrix} = \begin{bmatrix} \frac{EA_1}{L} \frac{(r-1)}{\ln r} & & & & & u_1 \\ 0 & JC_{11} & & & & v_1 \\ 0 & J(C_{11} - LC_{11}) & J(L^2C_{11} - 2LC_{12} + C_{22}) & & & \theta_1 \\ -\frac{EA_1}{L} \frac{(r-1)}{\ln r} & 0 & 0 & \frac{EA_1}{L} \frac{(r-1)}{\ln r} & & u_1 \\ 0 & -JC_{11} & J(LC_{11} - C_{11}) & 0 & JC_{11} & v_1 \\ 0 & -JC_{12} & J(LC_{12} - C_{22}) & 0 & JC_{12} & \theta_1 \end{bmatrix} \begin{bmatrix} u_1 \\ v_1 \\ \theta_1 \\ u_1 \\ v_1 \\ \theta_1 \end{bmatrix}$$

(2.26)

This may be written as

$$\{F\} = [K]^e \{\delta\} \quad (2.27)$$

where  $[K]^e$  is the stiffness matrix of the beam element.

If  $I_1 = I_2$ ,  $A_1 = A_2$ , then  $P = 1$ ,  $C_{11} = 1$ ,  $C_{12} = L/2$ ,

$$C_{22} = \frac{L^2}{3}, \quad D = \frac{L^2}{12}, \quad \frac{r-1}{\ln r} = 1$$

Stiffness matrix of eccentric beam element:

The stiffness matrix of the beam element in (2.26) is derived relative to the axis of the beam element, but in practical structures there exists an eccentricity between the axis and the computational node if we use as nodal point the intersection point of the membrane, the long truss, and the frame. The effect of this eccentricity will be considered below.

Figure (2.5) shows a beam element with axial nodes 1 and 2. If the computational nodes are 1' and 2', they are also in the

xy plane with eccentricities  $e_1, e_2$  from the nodes 1 and 2. We denote by  $\{\delta'\}$  the displacement column vector of the computational nodes in the x'y' coordinate system.

$$\{\delta'\} = \{u'_1, v'_1, \theta'_1, u'_2, v'_2, \theta'_2\} \quad (2.28)$$

The displacement column vector of the axial nodes 1 and 2 in the xy plane is

$$\{\delta\} = \{u_1, v_1, \theta_1, u_2, v_2, \theta_2\} \quad (2.29)$$

Let us transform the stiffness matrix  $[K]^e$  from the xy coordinate system first to the x'y' coordinate system and then to the x'y' coordinate system. Relative to the x'y' system, we have

$$\begin{aligned} u_1 &= u''_1 + e_1 \theta''_1 \\ u_2 &= u''_2 + e_2 \theta''_2 \end{aligned} \quad \} \quad (2.30)$$

But

$$\begin{aligned} \theta''_1 &= \theta_1 & v''_1 &= v_1 \\ \theta''_2 &= \theta_2 & v''_2 &= v_2 \end{aligned} \quad \} \quad (2.31)$$

therefore

$$\begin{Bmatrix} u_1 \\ v_1 \\ \theta_1 \\ u_2 \\ v_2 \\ \theta_2 \end{Bmatrix} = \begin{Bmatrix} 1 & 0 & e_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & e_1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{Bmatrix} \begin{Bmatrix} u''_1 \\ v''_1 \\ \theta''_1 \\ u''_2 \\ v''_2 \\ \theta''_2 \end{Bmatrix}$$

or in simplified form  $\{\delta\} = [T_e] \{\delta''\}$

If we shift the force and torque acting at nodes 1 and 2 to nodes 1' and 2', as shown in Figure (2.6), then

$$\begin{aligned} M_{1'}^* &= e_1 F_1 + M_1 \\ M_{2'}^* &= e_2 F_2 + M_2 \end{aligned} \quad (2.33)$$

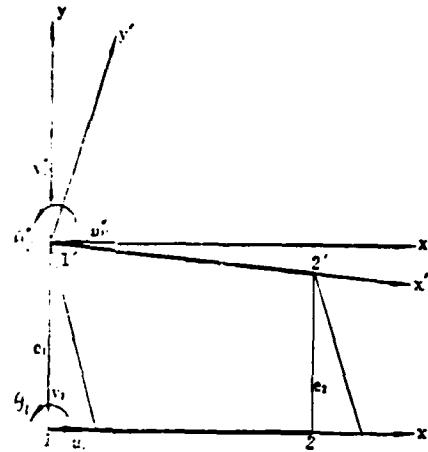


Figure 2.5

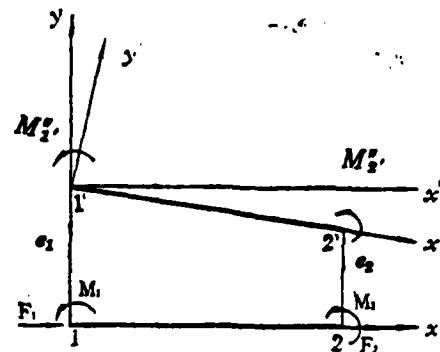


Figure 2.6

Then

$$\begin{pmatrix} F_{2'}^* \\ Q_{2'}^* \\ M_{2'}^* \\ F_{1'}^* \\ Q_{1'}^* \\ M_{1'}^* \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ e_2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e_1 & 0 & 1 \end{pmatrix} \begin{pmatrix} F_2 \\ Q_2 \\ M_2 \\ F_1 \\ Q_1 \\ M_1 \end{pmatrix} \quad (2.34)$$

or

$$\{F''\} = [T_e]^T \{F\} \quad (2.35)$$

Transformed to the  $x''y$  coordinate system, the stiffness matrix  $[K_e]^e$  becomes

$$[K_e]^e = [T_e]^T [K]^e [T_e] \quad (2.36)$$

Transforming  $[K_e]^e$  to the  $x'y'$  coordinate system, we have

$$\begin{aligned} \{\delta''\} &= [T_1] \{\delta'\} \\ \{F''\} &= [T_1] \{F'\} \end{aligned} \quad (2.37)$$

where

$$[T_1] = \begin{bmatrix} v_1 & v_2 & 0 & 0 & 0 & 0 \\ -v_2 & v_1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & v_1 & v_2 & 0 \\ 0 & 0 & 0 & -v_2 & v_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.38)$$

$$v_1 = \sqrt{1 - \left(\frac{e_1 - e_2}{L}\right)^2} \approx 1 - \frac{(e_1 - e_2)^2}{2L^2} \quad (2.39)$$

$$v_2 = \frac{e_1 - e_2}{L} \quad (2.40)$$

Therefore

$$[K']^e = [T_1]^T [T_1]^T [K] [T_1] [T_1] \quad (2.41)$$

$[K']^e$  is the stiffness matrix of the beam element after being transformed into the x'y' coordinate system. In general, L is much larger than  $e_1 - e_2$ , hence when  $e_1$  is about equal to  $e_2$ ,  $v_1 \approx 0$ ,  $v_2 \approx 1$ , then  $[T_1]$  is the unit matrix.

We further transform from the local coordinate system into the global coordinate system, assuming that the frame only bends in its own plane, while disregarding the ability of the membrane and the long beam to support bending. The displacement column vector in the global coordinate system is

$$\{\delta\} = \{u_2, \bar{v}_2, \bar{w}_2, \bar{\theta}_2, u_1, \bar{v}_1, \bar{w}_1, \bar{\theta}_1\} \quad (2.42)$$

Its transformation relation to  $\delta'$  is

$$\{\delta'\} = [T] \{\delta\} \quad (2.43)$$

where the transformation matrix is

$$[T] = \begin{bmatrix} [L] & 0 \\ 0 & [L] \end{bmatrix}$$

with

$$[L] = \begin{bmatrix} l_1 & m_1 & n_1 & 0 \\ l_2 & m_2 & n_2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (2.44)$$

After the coordinate transformation, the stiffness matrix in the global coordinate system with the eccentricity of the beam element taken into consideration is

$$[K]^* = [T]^* [T_1]^* [T_2]^* [K]^* [T_2] [T_1] [T] \quad (2.45)$$

In (2.44) the direction cosines are determined as follows: An arbitrary reference point 3' is chosen in the plane of the beam. Then the points 1', 2' and 3' are co-planar. The direction cosines in (2.44) may then be obtained by changing 1,2 and 3 in (2.11) - (2.14) to 1', 2' and 3'.

### III. Sample Computation and Results

We have carried out computations using the method described in this paper for the fuselage with cutout as shown in Figure (1.1) for the cases of symmetric loading and of side loading. The results are as follows:

(1) In the case of symmetric loading, the longitudinal displacement of the cross-sections on either side of the fuselage cutout is basically in agreement with the planar assumption. The result of the computation is shown in Figure (3.1).

(2) In the case of symmetric loading, the theoretically computed torsion of the fuselage and the static testing results are tabulated in Table (3.1). The curve is shown in Figure (3.2.)

(3) In the case of symmetric loading, the distribution of the axial force on the longitudinal truss of the fuselage is shown in Figure (3.3).

From Figure (3.1), (3.2) and (3.3) we see that under symmetric loading the computational results fit the loading characteristics of the fuselage better when it bends vertically.

(4) In the case of side loading, besides bending moment, the rear section of the fuselage also has a torsional moment  $m_z$ . Here there is torsional buckling in the axial displacement on both sides of the symmetric plane of the fuselage. The results are shown in Figure (3.4).

(5) Under torsional moment, the distribution of the normal stress at the cutout (between frame 43 and frame 45) with the effect of the frame elasticity taken into consideration is shown in Figure (3.5). This agrees completely with the testing results of an open tube under torsion.

## REFERENCES

- [1] J. S. Puchiminschy [as phoneticized], translated by Wang Derong et al. Matrix structural analysis. Defense Industry Publishing House (1974).
- [2] Proceedings of the second Conference on Matrix Methods of structural Mechanics, AFFDL-TR-68-150, Air Force Institute of techn. oct. 1968.
- [3] MAGIC: An Automated General Purpose System For structural Analysis, volume 1: Engineer's manual, AFFDL-TR-68-56.
- [4] Finite element methods and their application. Institute of Mathematical Technology, Chinese Academy of Sciences, 1975, No. 12.
- [5] Cao Zhihao, Direct analysis method of large scale linear algebraic equations. Published by Futan Journal of Natural Sciences, 1974, No. 1.
- [6] Argyris J. H. and Kelsey S. The Analysis of Fuselages of Arbitrary Cross section and Taper Aircraft Engineering XXX 1 N361-367. 1959.

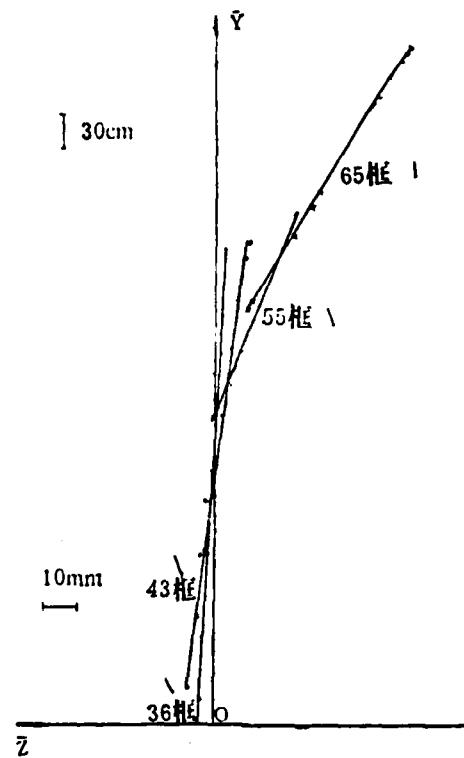


Figure 3.1. Longitudinal Displacement of Fuselage Cross-section under Symmetric Loading.  
 1 - frame

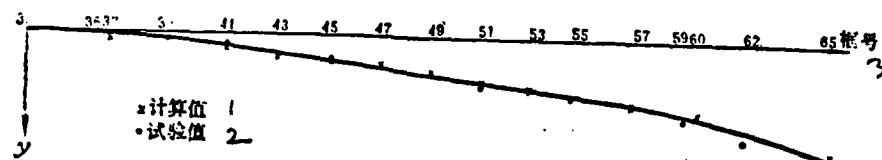


Figure 3.2. Torsion of Fuselage under Symmetric Loading.  
 1 - calculated value; 2 - experimental value.  
 3 - frame number.

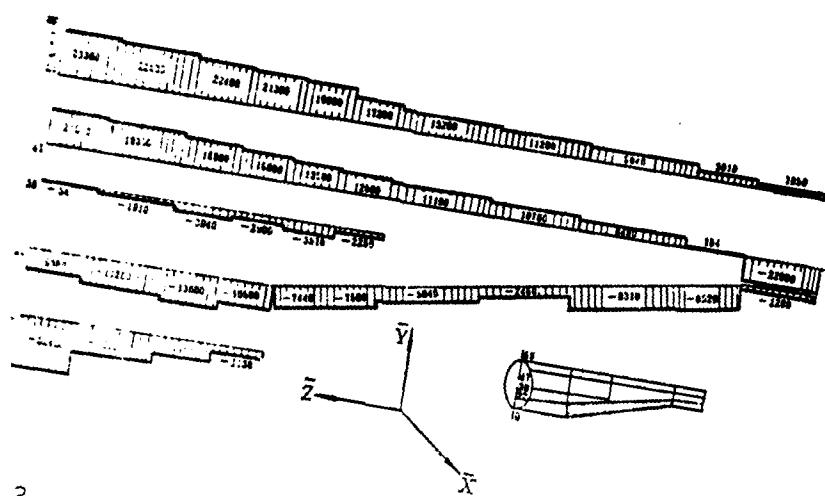


Figure 3.3. Axial Force of Long Truss under Symmetric Loading.

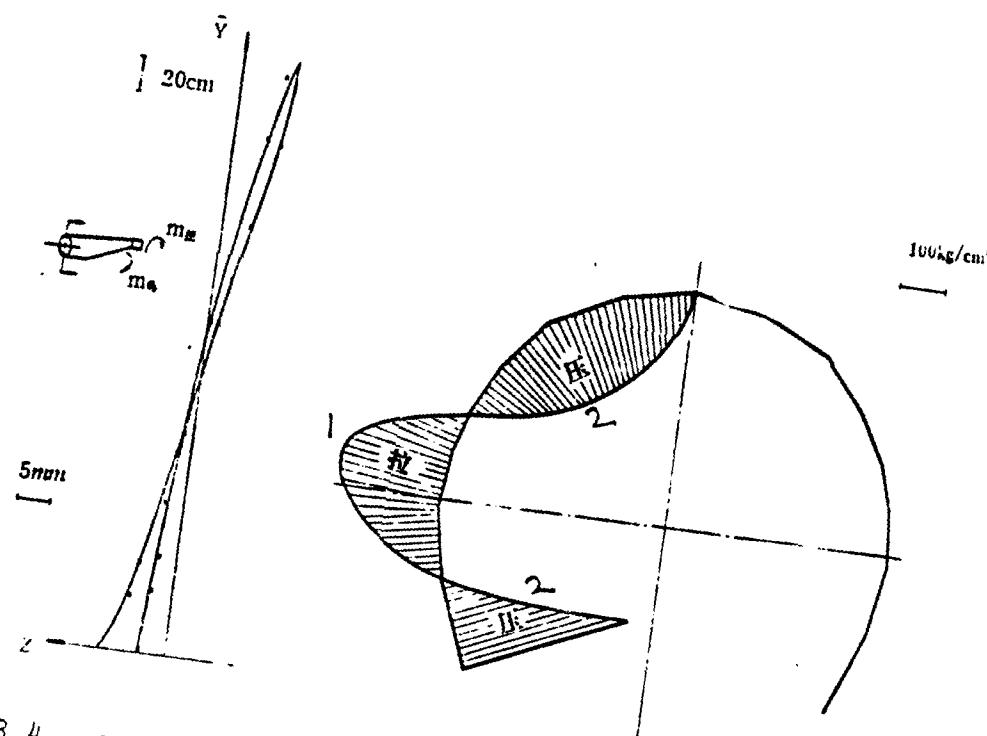


Figure 3.4. Axial Displacement of Frame 43 under side Loading.

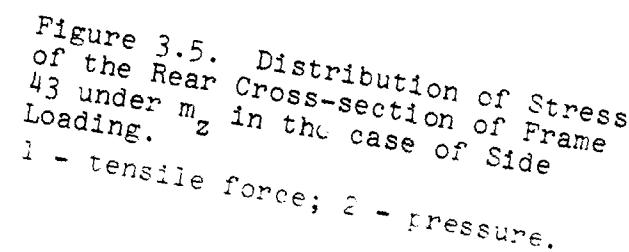


Figure 3.5. Distribution of Stress of the Rear Cross-section of Frame 43 under  $m_z$  in the case of Side Loading.

1 - tensile force; 2 - pressure.

Table 3.1. Frame 33 Rotation Angle

Frame No.	33	36	37	39	41	43	45	47	49
100% $P^4$	4.13		6.88		8.76	11.3	11.9	13.4	15.5
$y_t - y_{33}$	0		2.75		4.62	7.17	7.77	9.27	11.37
$L_t$	0		175		375	475	675	675	775
$y_m = y_t - y_{33} - \theta_{33} L_t$	0		1.33		1.57	3.31	3.09	3.78	5.07
$y_{calc}$	0.146		0.78	1.43	2.58	2.92	4.24		
Frame No.	51	53	55	57	59	60	62	65	
100% $P$	18.8	19.4		23.3		24.4		32.1	
$y_t - y_{33}$	12.67	15.27		19.17		20.27		27.97	
$L_t$	875	975		1145		1275		1495	
$y_m = y_t - y_{33} - \theta_{33} L_t$	5.55	7.34		9.85		9.90		15.81	
$y_{calc}$	7.54		9.49		11.10		14.4	16.91	

$y_m = y_{measured}$

END  
THE  
FARMED  
181  
DEMO

AD-A104 327

FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OH  
RECENT SELECTED PAPERS OF NORTHWESTERN POLYTECHNICAL UNIVERSITY—ETC(U)

F/6 20/4

AUG 81

FTD-ID(RS)T-0259-81-PT-1

NL

UNCLASSIFIED

5-5

AD-A  
104-327



END

DATE  
FILED

2-82

1  
DTIC

# **SUPPLEMENTARY**

# **INFORMATION**

ERRATA

AD-A104 327

Page 215 not available

DTIC-DDA-2  
16 Feb 82